Titre: Title:	Prédiction de la turbidité à l'eau brute et à l'eau filtrée de la Ville de Montréal à l'aide de réseaux de neurones
Auteur: Author:	Pierre Grimaud
Date:	2007
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Grimaud, P. (2007). Prédiction de la turbidité à l'eau brute et à l'eau filtrée de la Ville de Montréal à l'aide de réseaux de neurones [Master's thesis, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/8085/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/8085/
Directeurs de recherche: Advisors:	
Programme: Program:	Unspecified

UNIVERSITÉ DE MONTRÉAL

PRÉDICTION DE LA TURBIDITÉ À L'EAU BRUTE ET À L'EAU FILTRÉE DE LA VILLE DE MONTRÉAL À L'AIDE DE RÉSEAUX DE NEURONES

PIERRE GRIMAUD
DÉPARTEMENT DES GÉNIES CIVILS, GÉOLOGIQUES ET DES MINES
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE CIVIL)
DÉCEMBRE 2007

Your file Votre référence
ISBN: 978-0-494-36916-6
Our file Notre référence
ISBN: 978-0-494-36916-6

UNIVERSITÉ DE MONTRÉAL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

PRÉDICTION DE LA TURBIDITÉ À L'EAU BRUTE ET À L'EAU FILTRÉE DE LA VILLE DE MONTRÉAL À L'AIDE DE RÉSEAUX DE NEURONES

présenté par : GRIMAUD Pierre
en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de :

M. BARBEAU Benoit, Ing., Ph.D., président
Mme PRÉVOST Michèle, Ph.D., membre et directeur de recherche
M. LABIB Richard, Ph.D., membre et codirecteur
M. ESSEMIANI Karim, Ph.D., membre

À tous les arbres qui ont permis de réaliser cette maîtrise...

REMERCIEMENTS

Je tiens à remercier Michèle Prévost pour son dynamisme et son enthousiasme, grâce auxquels je me suis inscris en maîtrise à la Chaire industrielle CRSNG en Eau Potable huit mois avant d'entreprendre les démarches d'immigration. Merci également aux titulaires de la Chaire (Michèle Prévost, Benoît Barbeau et Raymond Desjardins) pour tous leurs efforts, efforts permettant à la Chaire d'être ce qu'elle est aujourd'hui. Des remerciements pour M. Comeau dont les cours d'épuration biologique et de design furent « une révélation » en matière d'ingénierie environnementale.

Des remerciements à M. Labib pour le soutien technique et son encadrement en tant que codirecteur de ce projet. Notamment les éclaircissements sur le monde obscur des réseaux de neurones, mais aussi les nombreuses corrections apportées au mémoire.

Merci également à Jacinthe Mailly, Julie Philibert et Claude Desroches pour les nombreuses assistances informatiques qui ont ponctué ces les deux années du projet…

Un merci, mais surtout mes amitiés pour les autres étudiants. Pour nos vénérables anciens de feu le « bureau des gars » désormais dans la vie active : Kenza, Tahiti Dave, Rickie Steamboat, Romain (x2), Gabriel notre futur ministre, etc. à bientôt. Pour les amis du « bureau des filles » avec qui être 'papoteux' fut un plaisir, bonne continuation dans vos 'doc' ou dans vos firmes d'ingénieries pour Ana, Cindy, Isa, et Guillaume. Une mention spéciale pour Daniel qui avait toujours une bonne anecdote ou expérience à conter. Sans oublier le bureau des 'doc' et nos successeurs (Françoise, Arash, Normand, Murielle, Elise, Tyler, Amine, Annie…) ainsi que les gens d'eaux usées (Carlos, Majdala, Bertrand…) pour les discussions à l'heure du dîner et activités connexes.

Merci à mon entourage et ma famille pour leur soutien, aux gens sympathiques du « *Composite Lab* » de Mc Gill. À Loleï pour la salsa, le brassage de notre bière et l'introduction à tes charmantes collègues de travail… Mais surtout à mon Acadienne préférée avec qui courir, aller boire des verres aux Dieux du Ciel, courir en bobettes, partager l'appartement, aller grimper dans les Adirondacks est toujours un bonheur.

RÉSUMÉ

La ville de Montréal puise son eau brute dans le fleuve St-Laurent au niveau de Ville Lasalle. La qualité de l'eau brute en termes de turbidité et de contamination fécale, place cette source dans la catégorie d'excellente qualité ('Bin 1' de l'USEPA) nécessitant un traitement minimal. Toutefois, la source est sujette à des augmentations significatives de turbidité de courte durée au printemps et à l'automne. La filière actuelle comporte des filtres à sable non assistés chimiquement dont la sortie est influencée par la turbidité à l'eau brute. Une autre option de traitement consiste à ajouter une étape de coagulation/floculation en amont des filtres (filtration directe). Dans ces deux cas, les fluctuations de turbidité à l'eau brute ont des répercussions directes sur l'ajustement des traitements et sur la qualité de l'eau filtrée. Afin de rencontrer les normes du RQEP, la Ville doit pouvoir assurer une valeur de turbidité inférieure à un certain seuil, soit 1 UTN en moyenne mobile mensuelle pour la filtration non assistée chimiquement. La présente étude consiste à fournir un outil de prédiction des valeurs de turbidité pour une journée à l'avance pour l'eau brute et l'eau filtrée, et ceci pour les deux usines de traitement (Atwater et Charles J. Des Baillets). L'utilisation de cet outil permettrait de mettre en place des moyens préventifs pour minimiser l'impact des pointes de turbidité transitoires. Ainsi, par exemple, en cas de dépassement annoncé aux normes du RQEP, on pourrait considérer la diminution des charges superficielles de filtration, l'ajout de coagulants d'appoint, etc.

Les paramètres influençant la turbidité à l'eau brute de Montréal ont déjà été déterminés par des travaux antérieurs (Tremblay, 2004). Ce projet vise à bâtir un modèle de réseau de neurones artificiels (RNA) plus précis et plus robuste de prédiction de la turbidité à la prise d'eau de Des Baillets; puis, d'envisager son implantation en ligne afin de fournir un outil de gestion proactif des évènements turbides.

Les conclusions des travaux menés identifient comme causes explicatives majeures les variables de rapport de mélange Outaouais fleuve St-Laurent et de tempêtes de vent, respectivement pour les saisons printemps et automne. Un index de saison fut utilisé constamment en tant qu'entrée pertinente aux modèles développés.

Le présent projet se base sur une plus grande quantité de données disponibles (1996-2006 au lieu de 1998-2001), ainsi qu'un découpage en modèles saisonniers. De plus, le couplage des modèles de classification et de régression permet d'augmenter les performances accessibles. Les modèles de classification détectent des dépassements par rapport à des seuils de turbidité spécifiés, alors que le modèle de régression donne une information sur le sens de variation et l'intensité d'un évènement turbide. La création de ces sous modèles spécifiques à une saison et à une plage de turbidité fait accroître les performances. À l'aide des travaux antérieurs et d'une revue de littérature sur l'utilisation des réseaux de neurones en hydrologie, une méthodologie adaptée aux besoins de ce projet a été développée. Les chiffres données ci-après se basent sur un ensemble de données dites « Test », données n'ayant pas servi à la calibration des modèles.

Pour l'eau brute à l'usine Des Baillets, le modèle printanier montre au moins 91,1% de classification correcte pour le seuil 4 UTN, et un coefficient de corrélation de 0,92 pour le modèle régressif. À l'automne, les performances sont un peu moins bonnes mais atteignent tout de même 87,1% de classification correcte et une corrélation de 0,78.

Concernant l'eau brute à Atwater, elle est fortement corrélée avec la qualité de l'eau brute à Des Baillets deux jours auparavant. Des modèles régressifs linéaires saisonniers ont suffi à obtenir des corrélations de l'ordre de 0,93 et 0,84 respectivement pour le printemps et l'automne.

La turbidité en sortie des filtres variant lentement, un pas de temps journalier fut adopté. Sa valeur la veille s'avère être un bon indicateur de la traitabilité de l'eau. Couplé aux paramètres de qualité de l'eau brute, les corrélations à Des Baillets sont autour de 0,91 et 0,90, respectivement pour le printemps et l'automne. Pour Atwater, ces chiffres montent jusqu'à 0,97 et 0,94 respectivement.

ABSTRACT

The main drinking water intake for the City of Montréal feeding the Charles J. Des Baillets and Atwater water treatment plants is located in the Saint-Lawrence River channel, near Ville Lasalle. In terms of turbidity and faecal contamination, the raw water quality is considered to be excellent, fitting in the 'Bin1' of the USEPA long term II regulations. The needed credits for the removal of *Cryptosporidium* and *Giardia* can be achieved through a combination of disinfectants including ozonation, chlorination, and of UV inactivation. The current Quebec regulations allow for non chemically assisted filtration, if no disinfection credit is granted for this process. This waiver for the use of coagulation prior to filtration is conditional to a criteria of filter effluent water quality turbidity. The turbidity of each filter effluent must not exceed a running monthly average of 1 NTU and a maximum at all times of 5 NTU. Historical data has shown that the existing filters in the two large drinking water plants can achieve these criteria most of the time with a few exceptions during events of very high raw water turbidities. In fact, during spring and fall, the source water undergoes short term but significant increases in turbidity. Because the plants are currently relying on non-chemically assisted sand filters (<5m/h), the filtered water is highly dependent on the raw water quality.

This work was undertaken to provide a better understanding of the occurrence and causes of turbidity in the source water and the filtered water. This project specifically proposes a forecasting tool for the prediction one day ahead of raw and filtered water turbidity for both treatment plants (Charles J. Des Baillets and Atwater). This model can assist operations in preparing for elevated turbidity at the intake. Possible courses of action include lowering the filtration rates, splitting flows between plants or adding a coagulants or filter aid. This predictive tool will help manage the impact of transient turbidity spikes.

Building on previous work, the goal of this project is to build a more precise and more robust model for forecasting the water turbidity at the plant's intake and to add a filtration module.

Up to now, two main variables have been identified for being responsible for spring and fall turbid spikes: respectively, the fraction between the Outaouais and the Saint-Lawrence rivers, and wind storm. À season index appeared to be a highly relevant variable in the developed models.

Consequently, this project is based on a wider database (ranging from 1996 to 2006 instead of from 1998 to 2001), and relies on seasonal models. Besides, using both classification and regression models increased forecasting performances. On one hand, classification models can detect if turbidity exceeds a certain threshold. On the other hand, regression models give information about the increase or decrease, and the magnitude of turbid events. Building such specialized sub models results in better predictions. With the help of former work and literature review on neural network in hydrology, a new methodology has been built according to the project needs. Numbers given below are taken from a dataset called 'Test', constituted form data which have never been presented to the model during the calibration phase.

First, concerning Des Baillets plant's raw water, the spring model shows at least 91.1% of correct classification for the threshold 4 NTU, and a correlation coefficient of 0.92 for the regression model. During fall, performances are a little bit lower. And yet, they reach 87.1 % of correct classification and a correlation of 0.78.

Second, for Atwater plant's raw water, a strong correlation is found with Des Baillets two days before. Regressive linear models are efficient enough to reach correlations of 0.93 and 0.84 for spring and fall respectively.

Finally, filtrated water turbidity shows slow variations, so, a daily time step has been adopted. The effluent filters' turbidity the day before appears to be a good predictor of the treatability of the water. Coupled with raw water quality variables, correlations as high as 0.91 and 0.90 have been found for Des Baillets spring and fall. For Atwater plant, these numbers raises to 0.97 and 0.94.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABBRÉVIATIONS

UTN	Unité de Turbidité Néphélométrique
RNA	Réseau de Neurone Artificiel
PMC	Perceptron Multicouches
MLP	Multilayer Perceptron
GRNN	Generalized Regression Neural Network
B	un scalaire
$\mathbf{B}$	un vecteur
$\underline{\underline{\mathbf{B}}}$	une matrice
$\Re^n$	espace des réels de dimension n
ARP	Algorithme de RétroPropagation
$\nabla E(\mathbf{w})$	gradient de $E(\mathbf{w})$
η	taux d'apprentissage
μ	momentum
$\frac{\partial E}{\partial \mathbf{w}}$	dérivée partielle de E par rapport à $\mathbf{w}$
ACP	Analyse en Composante Principale
SOM	Self Organizing Map
GA	Genetic Algorithm
EQM	Erreur Quadratique Moyenne
MES	Matière en Suspension
SRC	Sediment Rating Curve
R^2	Coefficient de détermination (ou de Nash Sutcliff)
DCO	Demande Chimique en Oxygène
UVA-254nm	Absorbance UV à 254 nm
MON	Matière Organique Naturelle
EAM	Erreur Absolue Moyenne

LISTE DES ANNEXES

INTRODUCTION

Une usine de production d'eau potable a pour fonction de traiter et distribuer une eau potable pour usage domestique (consommation, hygiène, sécurité, incendie...), ou industriel. Le traitement vise à retirer de l'eau toutes les substances pouvant présenter un risque pour la santé. De plus, l'eau produite doit satisfaire aux exigences esthétiques du consommateur. Le traitement est divisé en deux grandes étapes : un enlèvement physique ou chimique (enlèvement de la matière solide en suspension, de la couleur, des goûts et odeurs), et surtout une étape de désinfection visant à dégrader ou inactiver les pathogènes, micro-organismes et virus responsables de maladies hydriques.

Une bonne partie du traitement est assurée au moyen de l'enlèvement physique des particules. La filtration peut être réalisée sur un média filtrant (sable, anthracite...), ou plus récemment par filtration membranaire, et être chimiquement assistée ou non. Mesurer en temps réel les concentrations de tous les micro-organismes en entrée et en sortie de traitement serait physiquement irréalisable. Ainsi, un indicateur de la performance de la filtration est la mesure physique appelée turbidité (exprimée en Unité de Turbidité Néphélométrique, ou UTN). La turbidité est représentative du nombre de particules en suspension dans l'eau. Cette dernière est liée à l'intensité de la lumière diffusée et transmise par les particules en suspension dans l'eau lorsque celle-ci est éclairée par un faisceau incident. Les mesures peuvent donc être effectuées en temps réel, autorisant ainsi une surveillance constante de l'intégrité des systèmes filtrants.

En théorie, la turbidité dans l'eau ne représente pas de risque sanitaire, il s'agit juste d'un critère esthétique. Cependant, les micro-organismes responsables de maladies hydriques peuvent être liés aux particules solides par des mécanismes d'adsorption, d'occlusion, ou bien être eux-mêmes considérés comme particules selon leur taille. Ainsi, la turbidité est un indicateur de qualité de l'eau, sa valeur est donc normée.

Au Québec, le Règlement sur la Qualité de l'Eau Potable (RQEP, 2005) fixe les exigences de traitement en fonction de la qualité de l'eau brute. Concernant la filtration, le RQEP oblige à mesurer la turbidité en aval de chaque unité de filtration de l'usine. Dans le cas d'une eau coagulée, filtrée et désinfectée, la turbidité ne doit pas dépasser 0,5 UTN dans plus de 5 % des mesures mensuelles effectuées toutes les 4 heures. Si la filtration n'est pas assistée par coagulation, cette limite monte à 1 UTN en moyenne mobile dans les mêmes conditions de mesure. Dans tout les cas, la turbidité de l'eau filtrée doit être inférieure ou égale à 5 UTN en tout temps. Le non respect de ces normes peut entraîner des avis de bouillir auprès des consommateurs, ou bien le rejet à l'égout de l'eau filtrée tant que l'eau produite ne satisfait pas aux exigences du RQEP. Ces situations peuvent présenter un risque sanitaire et constitue un coût pour l'exploitant.

La ville de Montréal dispose de deux usines de filtration : Charles J. Des Baillets et Atwater, située en aval du canal éponyme. Toutes les deux puisent leur eau brute dans le fleuve Saint-Laurent. Cette source est qualifiée d'excellente qualité (« *Bin 1* ») selon les critères de l'US Environmental Protection Agency (USEPA) et nécessite un traitement minimal. Les deux usines opèrent des filtres à sable non assistés chimiquement. Le non recours aux coagulants chimiques rend ces filtres sensibles aux variations de turbidité de l'eau brute. Même si la turbidité à l'eau brute varie peu (en général elle est comprise entre 0 et 4 UTN pour l'usine Des Baillets), elle est néanmoins victime de hausses occasionnelles. Ces brusques hausses sont observées surtout au printemps et à l'automne et peuvent atteindre des valeurs aussi « hautes » que 30 UTN et 16 UTN respectivement pour les usines Des Baillets et Atwater.

Il serait particulièrement utile de pouvoir anticiper ces brusques augmentations de turbidité afin de pouvoir réagir de manière proactive à ces situations pouvant être problématiques. Ce projet fait suite à une étude précédente (Tremblay, 2004) qui identifia les évènements responsables de ces pointes de turbidité, et qui bâtit un

modèle prédictif de ces augmentations à l'usine Des Baillets. L'outil de modélisation utilisé fut les Réseaux de Neurones Artificiels (RNA). Cet outil statistique de modélisation basé sur les données connaît une utilisation croissante ces dernières années. Le présent projet doit bâtir quatre modèles prédictifs pour :

- la turbidité à l'eau brute de l'usine Des Baillets. Ce modèle doit être plus précis et plus robuste que celui développé précédemment.
- la turbidité à l'eau filtrée pour l'usine Atwater.
- la turbidité à l'eau filtrée pour l'usine Des Baillets.
- la turbidité à l'eau filtrée mixte pour l'usine Atwater.

Ces modèles permettront de connaître une journée à l'avance la valeur de la turbidité à l'eau brute et en sortie des filtres. Une gestion proactive des situations problématiques pourra être menée avec la mise en place de solutions temporaires si un pic est prédit.

Après une rapide mise en contexte, le premier chapitre vise à rappeler les travaux antérieurs et à documenter les articles jugés pertinents lors de la revue de littérature sur l'utilisation des réseaux de neurones en hydrologie. De là découlera une méthodologie qui sera décrite dans le deuxième chapitre. Cette méthodologie servira de cadre d'étude pour les autres parties du mémoire. Les chapitres 3, 4, 5 et 6 traiteront en détail de chaque modèle prédictif élaboré. Chacun de ces chapitres sera introduit par une rapide mise en contexte, un descriptif de la méthode utilisée et les résultats attenants. Des commentaires et discussions viendront clore ces chapitres. Finalement, une conclusion viendra commenter les travaux effectués et leurs contributions.

Chapitre 1 REVUE DE LITTÉRATURE

Le but de ce projet est de bâtir un modèle prédictif de la turbidité à la prise d'eau brute et à l'eau filtrée des stations de filtration de la ville de Montréal, à savoir Charles J. Des Baillets et Atwater. Ces deux stations puisent leur eau brute dans le fleuve Saint-Laurent à Ville Lasalle, juste en aval du lac Saint-Louis. La filière actuelle comporte des filtres à sable non assistés chimiquement dont la sortie est influencée par la turbidité à l'eau brute.

Les paramètres influençant la turbidité à la prise d'eau de la ville de Montréal ayant été déterminés par des travaux antérieurs (Tremblay, 2004), l'objectif est de bâtir un modèle de réseau de neurones artificiels (RNA) plus précis et plus robuste de prédiction de la turbidité à la prise d'eau; puis, d'envisager son implantation en ligne afin de fournir un outil de gestion proactif des évènements turbides.

Dans une première partie, la situation hydro-géographique de la zone d'étude sera récapitulée. Ensuite, les causes potentielles des évènements turbides tels qu'identifiés par les travaux antérieurs seront rappelées. Et finalement, la troisième section de cette revue de littérature sera consacrée à l'outil statistique considéré, à savoir les réseaux de neurones artificiels (RNA).

1.1 Mise en contexte

La présente section est directement extraite du chapitre 1 du mémoire de maîtrise de Tremblay (2004).

> « *La prise d'eau de la Ville de Montréal est située à 610 m de la rive nord du fleuve, à LaSalle, en amont des rapides de Lachine et en aval du lac St-Louis. Le lac Saint-Louis est formé par un élargissement naturel du fleuve à sa confluence avec la rivière des Outaouais. Il a une forme triangulaire, étant constitué à la base par les principaux affluents, et au sommet, par l'exutoire*

via les rapides de Lachine. Il couvre une superficie de 148 km^2, soit environ 23 km de longueur par 10 km de largeur dans ses plus grande dimensions (Centre Saint-Laurent, 1993 et Fortin et al. 1994).

Le lac Saint-Louis reçoit les eaux du lac Saint-François, qui s'écoulent en grande partie par le canal de Beauharnois (84% du débit en moyenne) le long de la rive sud et par le lit naturel du fleuve en bordure de la rive nord (Fortin et al. 1994). Cette section du fleuve reçoit aussi les eaux de la rivière des Outaouais par l'intermédiaire du lac des Deux Montagnes qui se décharge en partie dans le lac Saint-Louis par le canal de Vaudreuil, à l'ouest de l'île Perrot, et par le canal Sainte-Anne, au nord de l'île Perrot (Figure 1-1). Deux tributaires de moindre importance, les rivières Châteauguay et Saint-Louis débouchent le long de la rive sud du lac (Fortin et al., 1994). »

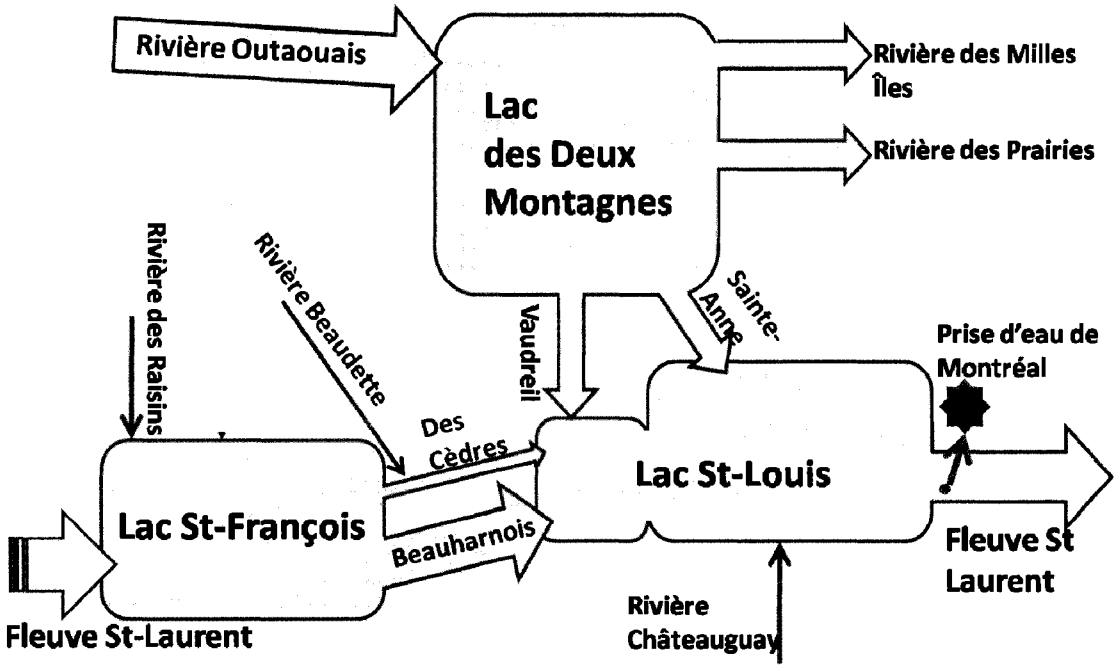

Figure 1-1 : Schéma du réseau hydraulique de la région de Montréal

Note : une version plus récente du rapport « Synthèse des connaissances sur les aspects physiques et chimiques de l'eau et des sédiments du lac Saint-Louis » (Fortin et *al.*, 1994) est disponible au Centre Saint-Laurent (Lorrain et al., 1999).

1.2 Travaux antérieurs

Cette section est principalement inspirée par les travaux antérieurs (Tremblay, 2004), que cela soit pour la revue de littérature ou les résultats de l'analyse des causes de la turbidité à l'eau brute.

1.2.1 Analyse descriptive de la turbidité

Le tracé des valeurs de la turbidité à l'eau brute en fonction du temps pour les trois années et demi de données superposées (de janvier 1998 à mars 2001) a permis d'identifier la présence de comportements saisonniers. Le printemps et l'automne se distinguent par une moyenne et une variabilité beaucoup plus forte qu'à l'été et à l'hiver, ainsi que par la présence de pointes de turbidité d'une durée plus ou moins longue. Le printemps s'illustre par une prédominance de pointes durant plusieurs jours, démontrant une dégradation de la qualité globale de l'eau. Ces pointes de turbidité sont parfois superposées à d'autres pointes de plus courte durée. La turbidité à l'automne semble être sous influence de pointes de forte amplitude et de courte durée. Les dates identifiées graphiquement pour le printemps et l'automne furent respectivement du 1er mars au 30 avril, et du 15 octobre au 14 janvier.

Une analyse statistique de la turbidité a permis de définir la notion d'évènement dit « turbide » si la valeur considérée de la turbidité excède le 90^{e} centile, soit 3,2 UTN. Concernant la notion temporelle, un évènement turbide est défini « de fond » lorsque sa durée dépasse cinq jours. Si un pic de turbidité a lieu pendant un évènement de fond, il est appelé « superposé ». L'analyse statistique des quantiles de la variables de turbidité donne naissance à une discrétisation de la turbidité en cinq intervalles

(classes I à V). Les frontières de ces intervalles sont délimitées par les centiles suivants : 75^{e}, 90^{e}, 95^{e}, et 99^{e}.

1.2.2 Détermination des causes potentielles de la turbidité

Différents apports au système

En isolant le système {gire de l'île Perrot, lac Saint-Louis, prise d'eau à Lasalle}, un bilan de masse sédimentologique (Lorrain et al., 1999) permet d'émettre la distinction entre deux types d'apports de matières en suspension : les apports internes et externes au système.

Parmi les apports externes, il y a tout d'abord la qualité de l'eau arrivant en amont des tributaires principaux (Figure 1-1): soit le fleuve Saint-Laurent par les chenaux Beauharnois et Des Cèdres, et l'eau en provenance du lac des Deux Montagnes (par Vaudreuil et Sainte-Anne de Bellevue).

Aux apports externes s'ajoutent aussi les contributions potentielles des tributaires secondaires comme les rivières Châteauguay ou Saint-Louis, et des ruisseaux sur l'île de Montréal par lesquels s'écoulent les eaux de drainage et de pluie de secteurs résidentiels ou industriels autour des communes de Dorval, Pointe Claire, Kirkland et Beaconsfield (Lorrain et al., 1999). Il s'agit des ruisseaux Bouchard, Denis, et Saint James. Ces ruisseaux sont trop petits pour être visibles sur les plans d'ensemble de l'île. Lors d'épisodes de fortes pluies, la capacité de captage du réseau d'égout peut se trouver dépassée et ceci donne lieu à des épisodes de surverses. En complément de ces eaux de drainage, nous pouvons ajouter le lessivage des terres lors de fortes précipitations.

De plus, la hausse des débits lors de crues printanières et automnales, ou bien à cause de fortes précipitations, pourrait être responsable de l'érosion des berges, et dans une moindre mesure du lit des cours d'eau. Le débit du fleuve Saint-Laurent est régularisé

par une série de barrages depuis les Grands Lacs. Ses fluctuations s'étendent ainsi sur plusieurs mois, si bien que la crue printanière a peu d'effet sur ses débits, contrairement aux autres cours d'eau. Par contre, le débit de la rivière des Outaouais, qui est régulé au barrage de Carillon, peut voir ses valeurs tripler en période de crue (Tremblay, 2004).

Parmi les apports dits internes, la cause principale identifiée est la remise en suspension des sédiments des lacs de faibles profondeurs d'eau, en cas de variations brusques de débits ou de tempêtes de vents.

Huit phénomènes explicatifs retenus

Du découpage précédent, il ressort huit phénomènes explicatifs potentiels des évènements turbides. Il s'agit :

1. De la qualité de l'eau en amont et les jours précédents : si en entrée du système, l'eau en provenance des Grands Lacs ou des Outaouais est de mauvaise qualité, la prise d'eau de Montréal sera aussi affectée.
2. Des fortes précipitations : comme vu précédemment, elles affectent le lessivage des terres, des réseaux de drainage urbains en cas de surverse, et elles peuvent causer l'augmentation des débits des tributaires secondaires. Ces variables semblent s'être effacées au profit d'autres variables véhiculant plus d'informations lors de la sélection d'entrées optimales pour les modèles bâtis par Tremblay (2004).
3. De la hausse des débits des tributaires secondaires : c'est une conséquence directe et facilement mesurable de la fonte des neiges ou des précipitations abondantes. La hausse des débits est responsable de l'érosion des berges et, dans une moindre mesure, du lit des cours d'eau. Une étude basée sur un bilan de masse des sédiments entre Cornwall et Québec montre que 65% des apports en termes de sédiments seraient dûs à l'érosion des berges (Rondeau et al.,

2000). Cette même étude identifie le secteur du canal de Beauharnois (juste en amont du lac Saint-Louis) comme une zone d'érosion majeure.

4. Des tempêtes de vent sont responsables de la création de vagues. Ces vagues érodent les berges ou bien le lit des lacs fluviaux dans certaines conditions spécifiques :

   « *Les zones où la hauteur d'eau est inférieure à 4,5 m sont particulièrement vulnérables à l'automne et au début de l'hiver, après la disparition des plantes aquatiques et des grands herbiers de macrophytes (Loiselle et al. 1997)* » (Tremblay, 2004).

   Des études du Centre Saint-Laurent, documentées par Tremblay (2004), montrent qu'en raison de leurs faibles profondeurs moyennes et de la faible vitesse des courants en dehors du chenal de navigation (<0,3m/s), les lacs Saint-François et Saint-Louis sont des zones d'accumulation permanente des sédiments qui peuvent être remis en suspension par la suite. Le tracé des cartes de bathymétrie (Figure 1-2) met en évidence que les lacs Saint-François et Saint-Louis peuvent être sujets à la remise en suspension des sédiments sous l'influence de tempêtes de vents, de part leurs faibles profondeurs moyennes. Ce facteur vent a eu une grande importance dans les entrées finales des modèles développés par Tremblay (2004).

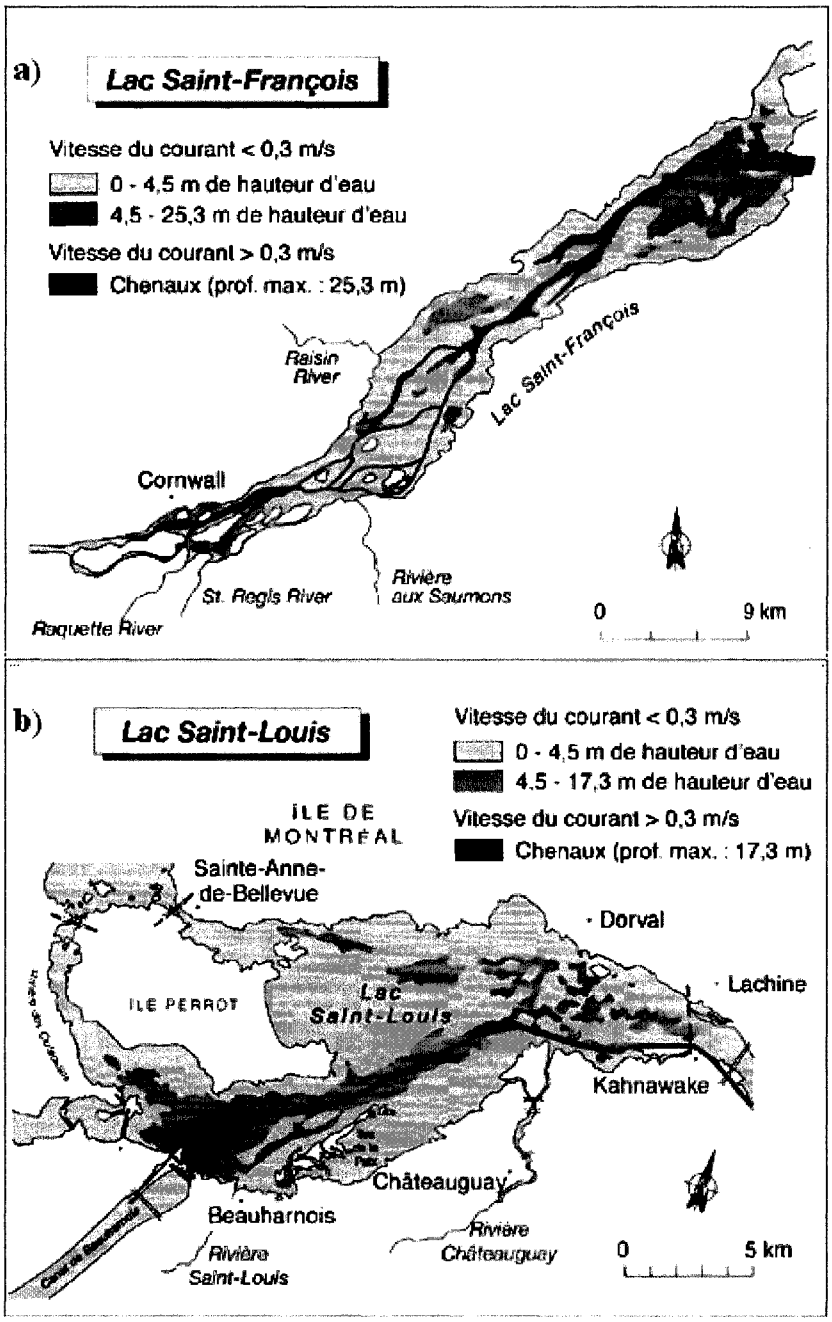

Figure 1-2 : Carte bathymétrique des lacs (a) Saint-François. (b) Saint-Louis[1]

[1] Source : Centre Saint-Laurent. http://www.qc.ec.gc.ca/csl/inf/inf019_002_f.html (page consultée le 7/09/2007)

5. Du renversement des lacs : le renversement est un phénomène mettant en mouvement les masses d'eau d'un lac. Sous réserve que le lac soit stratifié, lorsque la température de l'eau atteint 4°C (température du maximum de densité de l'eau), il se produit une circulation de l'eau des couches supérieures vers les couches inférieures. Cette circulation peut être responsable de la remise en suspension des sédiments accumulés dans les couches inférieures, elle peut se produire au printemps et à l'automne. D'autres détails de définition du renversement sont donnés à l'Annexe A. À cause de leurs faibles profondeurs (Figure 1-2), la circulation des courants s'y opérant, et la présence d'herbiers aquatiques au fond, les lacs Saint-François et Saint-Louis sont définis par le Centre Saint-Laurent en tant que lacs fluviaux, ils ne sont pas sujets à la stratification thermique, donc au renversement [(Champoux et Sloterdijk, 1988) cités par (Tremblay, 2004)]. Cependant, le régime du lac des Deux Montagnes est celui se rapprochant le plus d'un régime lacustre. Les baies, avec leurs faibles vitesses de courants accumulent les sédiments, et une vaste zone dans la partie centrale du lac est aussi une zone d'accumulation permanente de sédiments. Un renversement remettant en suspension les sédiments y est donc possible (Tremblay, 2004). La suite des travaux de Tremblay montre que ce phénomène, initialement supposé comme principale explication des pointes de turbidité, n'est qu'un facteur secondaire.
6. Du rapport de mélange des masses d'eau Outaouais / fleuve Saint-Laurent : une analyse des mélanges de masses d'eau par Hydro-Québec en 1985 conclut que la prise d'eau de la ville de Montréal n'est pas sous influence directe de la rivière Châteauguay, de même qu'elle n'est pas influencée par l'eau des Outaouais passant par le chenal de Sainte-Anne de Bellevue (car elle est située suffisamment loin de la rive). Le rapport conclut aussi que l'eau de mélange passant par Vaudreuil directement dans la gire de l'île Perrot est aussi sans conséquence pour la ville de Montréal comme l'attestent les Figures 1-6 et 1-7

du rapport de Tremblay. Par soucis de compréhension, le schéma de mélange des masses d'eau en cas de forte crue (>14000m^3/d) est rappelé en Figure 1-3. Cependant, les travaux antérieurs semblent montrer une influence notable de l'eau des Outaouais sur la qualité à la prise d'eau de Montréal :

« *L'accroissement des débits de la rivière des Outaouais et du fleuve St-Laurent au printemps dégrade la qualité de l'eau puisée à la prise d'eau de la Ville de Montréal, particulièrement au printemps. Frenette et Frenette (1992) ont déterminé les périodes sédimentologiques actives du fleuve et de ses tributaires à l'aide de sédimentogrammes. On remarque ainsi une pointe sédimentologique lors des crues printanières des tributaires (avril-mai) et une seconde pointe, de moindre importance, lors des crues automnales (octobre-novembre). Entre ces maxima, des pointes tertiaires associées aux différentes averses d'été et d'automne (juin-septembre) apparaissent, tandis que les charges solides d'hiver demeurent très faibles (décembre-mars). En tout, la période de crue printanière serait responsable de 60% et 70% de la charge sédimentaire.*

C'est la charge sédimentaire de la rivière des Outaouais qui contribue le plus significativement à l'augmentation des apports en matières solides au lac St-Louis, particulièrement au printemps car les variations saisonnières des concentrations de particules en suspension sont aussi beaucoup plus prononcées dans la rivière des Outaouais que dans le fleuve St-Laurent (SCN-Procéan, 1992). » (Tremblay, 2004).

Du fait des vitesses des courants supérieures à 0,3 m/s et de sa forte hauteur d'eau (voir Figure 1-2 b), le chenal de Beauharnois représenterait une zone où les sédiments s'accumulent peu et pourraient être transportés jusqu'à la prise d'eau. L'hypothèse est que l'eau de mélange issue de la gire ait un impact direct sur la prise d'eau, contrairement aux conclusions du rapport de 1985.

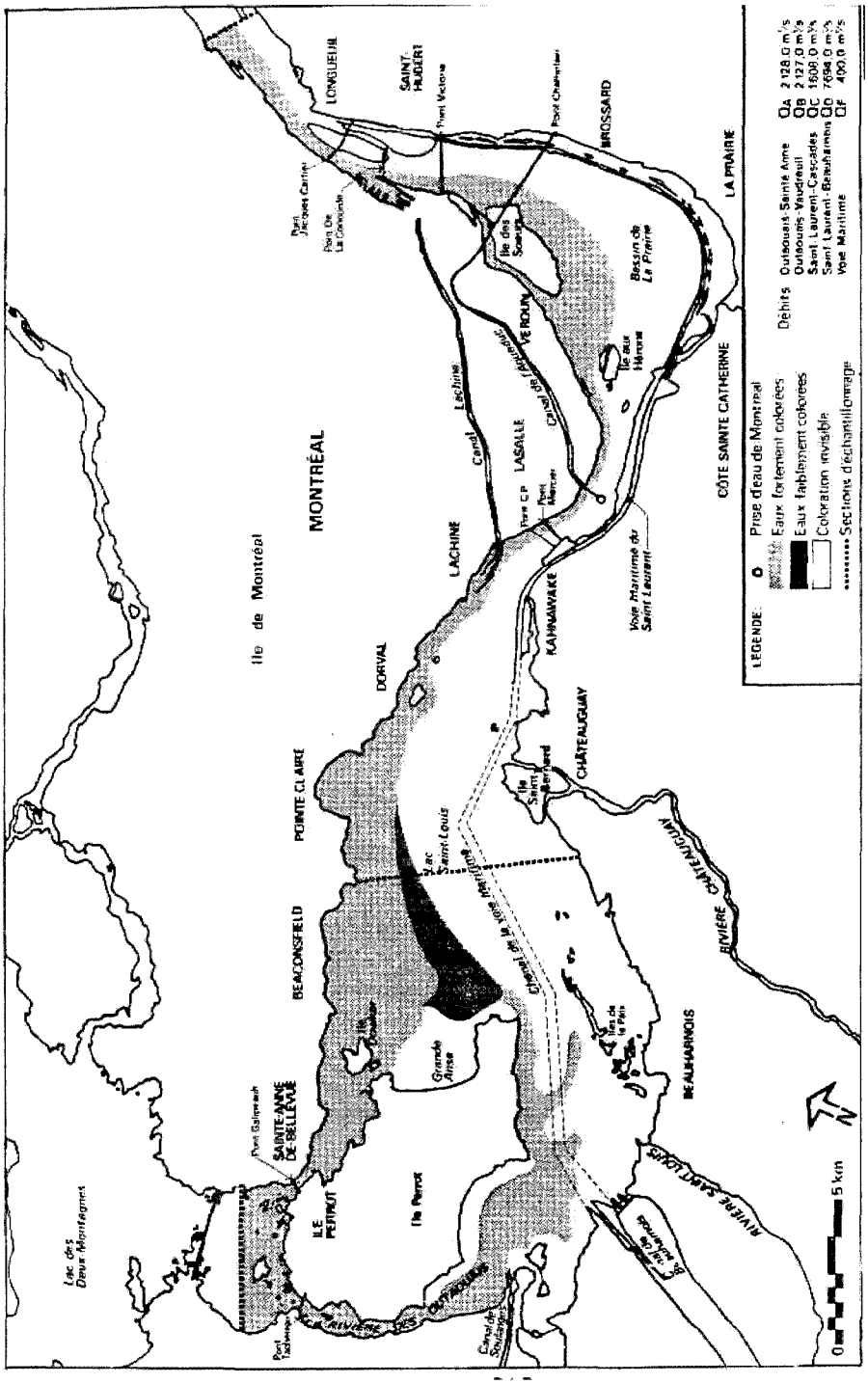

Figure 1-3 : Mélange des eaux au débit d'une forte crue de 14 000 m^3/s (Source : Hydro-Québec, 1985a)

7. De la fonte des neiges et de la crue printanière qui a un double impact : celui du lessivage des terres et des conduites de drainage des eaux de pluie, mais surtout, le plus problématique, l'augmentation des eaux de l'Outaouais par le barrage de Carillon. Ceci a pour conséquence d'augmenter significativement le rapport de mélange entre le fleuve Saint-Laurent et la rivière des Outaouais, de bien moins bonne qualité que le Saint-Laurent.
8. De la fragilité ou du bris du couvert de glace : après avoir montré la forte influence des tempêtes de vents sur les pointes de turbidité à la prise d'eau brute de Montréal, Tremblay (2004) remarque que la présence d'un couvert de glace durant l'hiver offre une protection naturelle aux tempêtes de vents et intempéries. Cependant, certaines pointes de chaleur hivernales couplées à des tempêtes peuvent être responsables du bris partiel de cette protection et de la remise en suspension des sédiments accumulés durant l'hiver sous ce couvert. Comme nous l'avons vu précédemment, la disparition de l'herbier aquatique rend la période automne-hiver particulièrement sensible à la remise en suspension des sédiments. Au printemps, la fonte des neiges entraîne le bris total de ce couvert. La création d'une variable indicatrice de la fonte des neiges permet de distinguer les épisodes de fragilisation du couvert et ceux de bris total (voir Annexe A).

Variables indicatrices : constitution d'une base de données préliminaire

Après avoir cité les facteurs potentiellement responsables des augmentations de turbidité à l'eau brute de Montréal, une liste de variables indicatrices de ces facteurs permet de fonder une base de données préliminaire à la création d'un modèle de prédiction des évènements turbides. Les connaissances préalablement acquises sur le sujet (revue de littérature et analyse descriptive des pointes de turbidité), couplées à des analyses statistiques permettent de réduire le nombre de variables en entrée des modèles.

1.2.3 Développement de deux modèles de prédiction par réseaux de neurones

Deux modèles de prévision furent bâtis, un modèle de classification et un modèle de régression. Le modèle de classification donnait en sortie les classes de turbidité identifiées lors de l'analyse par quantiles de la turbidité à l'eau brute, alors que le modèle de régression prédisait la valeur de la différence entre la turbidité du lendemain moins celle du jour. La mise en parallèle de ces deux modèles aboutit à la création d'un modèle dit fonctionnel. Ce dernier permit d'augmenter la capacité de prévision du modèle.

Au final, le modèle retenu de classification, prédisant la classe de turbidité le jour suivant, fut de type perceptron multicouches (PMC) 6 :5 :1. La notation 6 :5 :1 signifie que le modèle possède six entrées, cinq neurones dans la couche cachée, et un neurone en sortie du modèle. Les entrées retenues sont la valeur mesurée de la turbidité à Des Baillets le jour même, un index de saison, et quatre variables hydrométéorologiques mesurées la veille. Ces quatre variables hydrométéorologiques sont : le pourcentage rivière des Outaouais / fleuve Saint-Laurent, la vitesse du vent maximale horaire et moyenne horaire au lac Saint-François, et la vitesse du vent moyenne horaire à l'aéroport de Dorval. Ce modèle aboutit à un pourcentage de classification correcte sur tous les exemples de l'ordre de 83%.

Quant au modèle de régression, ce fut l'architecture GRNN 8 :740 :2 : 1 (*Generalized Regression Neural Networks*) qui fut retenue. Ce modèle contient 740 neurones dans la couche caché un, deux neurones dans la deuxième couche cachée, un neurone en sortie pour prédire la différence entre la valeur de demain moins celle du jour, et huit entrées. Ces entrées sont les valeurs mesurées le jour même pour : l'index de saison, la qualité de l'eau à Des Baillets (couleur, conductivité, et turbidité), et les trois variables météorologiques de vent citées précédemment en tant qu'entrées du modèle de classification. À ces sept entrées s'ajoute le pourcentage mesuré la veille du rapport de

débit de la rivière des Outaouais sur le débit du fleuve Saint-Laurent Ce modèle permit d'obtenir un coefficient de corrélation sur toutes les données égal à 0,84.

Le modèle fonctionnel permit d'atteindre un pourcentage de classification correcte des évènements turbides de l'ordre de 94,7% (donc 5,3% de faux négatifs). Cependant ce dernier accuse encore environ 15% de faux positifs! Ces fausses alertes pourraient dégrader significativement la confiance que l'opérateur pourrait avoir dans le modèle s'il devait être implanté en station.

Une des remarques de la discussion du mémoire de Tremblay (2004) vise à essayer de réduire ces faux positifs tout en conservant une prédiction des évènements turbides au moins aussi bonne. Pour ce faire, il est suggéré d'élaborer un modèle spécifique à chaque saison. En effet, l'analyse descriptive des évènements turbides met en évidence des formes distinctes sur les évènements turbides automnaux et printaniers, ces comportements peuvent être sous influence de facteurs explicatifs différents d'une saison à l'autre comme le montre l'inclusion systématique de la variable d'index de saison dans le choix final des entrées des modèles. De plus, un modèle bâti sur un plus grand nombre d'exemples verra ses capacités de généralisation accrues.

Ce projet portera donc sur l'élaboration d'un modèle prédictif de la turbidité à l'eau brute de Montréal plus précis et plus robuste. Ce modèle doit avoir une vocation opérationnelle, i.e. il doit pouvoir être implanté en station.

1.3 Les réseaux de neurones

Cette section sera découpée en trois volets. Le premier vise à rappeler les concepts fondamentaux sur les réseaux de neurones artificiels (RNA). Le deuxième donnera les étapes fondamentales nécessaires à la création d'un modèle par RNA, il sera illustré par les articles issus de la revue de littérature traitant de l'application des RNA en hydrologie. Finalement, le troisième et dernier volet observera plus en détails quelques

articles portant sur la prédiction de la qualité de l'eau brute, puis sur la prédiction de la performance de filtration assistée par coagulation.

1.3.1 Concepts fondamentaux

Le concept de neurone artificiel fut inventé dans les années 1940, inspiré par des recherches visant à modéliser le cerveau humain. Pendant des années, les RNA connurent peu d'engouement jusqu'au développement de nouveaux algorithmes d'apprentissage plus efficaces, algorithmes associés à une puissance de calcul toujours croissante. Dès lors les champs d'application des RNA s'ouvrèrent à des domaines aussi variés que l'informatique, la robotique, la finance, l'ingénierie, etc. Ainsi, les réseaux de neurones furent utilisés dans des problèmes de classification (attribution de prêts bancaires, reconnaissance de formes), de régression (approximer la relation existante entre un espace d'entrées et de sorties, modèles de prévision, complétion de bases de données à trous), d'organisation de bases de données (cartes auto-organisatrices et « *clustering* ») (Govindaraju, 2000a).

Les concepts essentiels récapitulés ci-après peuvent être consultés dans les ouvrages de référence (Bishop, 1995; Haykin, 1999).

Par souci de simplification, la sortie des modèles suivants (**Y**) est en dimension 1. Le cas de sortie multi variables ($\underline{\mathbf{Y}}$) se généralisant bien.

Modèle statistique VS conceptuel

Les RNA sont souvent considérés comme des modèles de type 'boîtes noires'. C'est-à-dire que contrairement aux modèles conceptuels où l'ensemble des équations physiques gouvernant le système est connu, les RNA n'effectuent qu'une approximation numérique de fonction, approximation basée sur une série de données mesurées. En effet, à partir d'une base de données de variables d'entrées $\underline{\mathbf{X}}$ et des mesures de la variable de sortie correspondante à modéliser (**Y**), le réseau neuronal

tente d'approcher numériquement la loi entrée/sortie $\mathbf{Y}=f(\underline{\mathbf{X}})$ liant les couples $(\underline{\mathbf{X}};\mathbf{Y})$. En théorie, aucune connaissance préalable n'est indispensable pour mettre en place un modèle RNA; cependant, en pratique, le modélisateur peut guider le réseau dans sa recherche d'un modèle par des choix judicieux et améliorer la performance accessible.

Les modèles RNA présentent des désavantages. Ce sont des modèles dépendants de données, c'est-à-dire qu'ils cherchent à extraire un patron à partir des exemples qui les alimentent. Les modèles RNA sont donc hautement dépendants des exemples qui leur sont fournis. La performance atteinte par un modèle neuronal est limitée par le bruit des données qui lui sont fournies. Effectivement, trois cas empêchent le réseau d'atteindre une prédiction parfaite : tout d'abord si les données exemples sont bruitées, si elles contiennent des informations n'ayant aucun lien avec la loi à modéliser (informations parasites non pertinentes), ou bien si elles ne contiennent pas toutes les variables explicatives de la loi physique à modéliser.

Autre inconvénient, compte tenu des données qui lui ont été fournies, le RNA convergera vers une équation liant entrées et sortie. Cette approximation numérique n'est normalement pas égale à la loi physique sous-jacente qu'un modèle conceptuel tente de modéliser. Effectivement, bien que les réseaux de type PMC puissent converger vers un minimum d'erreur de prédiction avec les données qui leur ont été fournies, les solutions obtenues peuvent être physiquement improbables (Kingston et al., 2005). Il peut s'agir par exemple de prédictions négatives pour la variable de turbidité.

Cependant, la mise en place d'un modèle conceptuel nécessite la connaissance complète des phénomènes et interactions physiques du système à modéliser. Un maillage accompagné d'une campagne extensive de mesures permettrait ensuite de calibrer le modèle conceptuel. Dans notre cas, il s'agirait de modéliser, à l'échelle du Saint-Laurent entre Beauharnois et la prise d'eau à Lasalle, les apports en sédiments par les cours d'eau, et la mécanique interne du fleuve dans le lac Saint-Louis

(écoulement du fleuve par éléments finis, interaction végétation aquatique sédiments, etc.). Comme l'avait conclu un article précédent sur la prévision des apports en phosphore à l'échelle d'un bassin versant (Nour et al., 2006a), cette approche conceptuelle (intensive en termes de données requises, de puissance de calcul, et par le manque de connaissances scientifiques sur tous les phénomènes engagés) semble évidemment économiquement inacceptable si l'objectif est d'implanter rapidement une solution fonctionnelle en station. Pour cette raison, les RNA constituent une alternative attrayante aux modèles conceptuels.

Spécificités des RNA

Par ailleurs, les RNA appartiennent à la catégorie des modèles semi-paramétriques (Dreyfus et al., 2004) : aucune hypothèse préalable n'est faite ni sur la distribution des données, ni sur la relation entrée/sortie que l'on souhaite modéliser, et ils sont construits par composition d'un nombre variable d'unités élémentaires dont l'équation est déjà prédéfinie. Ces unités élémentaires (ou neurones) sont interconnectées entre elles, comme l'illustre la Figure 1-4.

Comme nous le verrons dans la section suivante, les RNA sont construits par composition de fonctions non linéaires. Ils conviennent donc parfaitement pour la modélisation de phénomènes non linéaires (Haykin, 1999). De plus, les RNA sont relativement peu sensibles au bruit des données et peuvent fonctionner dans des domaines où le nombre de données disponibles est limité. Ils représentent donc un outil particulièrement adéquat en modélisation environnementale (Maier et Dandy, 2001; Recknagel, 2001).

Perceptron multicouches

Le type de réseau le plus répandu, le perceptron multicouche (PMC) de type « *feedforward* » est représenté sur la Figure 1-4 a. Chaque cercle représente une unité

élémentaire appelée neurone. Le réseau se décompose en trois couches. Le vecteur d'entrées **X** passe par la première couche où chaque variable d'entrée X_i va subir une première transformation pour donner la sortie X'_i. Ces sorties vont ensuite constituer les entrées des neurones de la couche suivante (couche cachée un). Et ainsi de suite jusqu'à la sortie du réseau (Y_1), le signal se propage de la gauche vers la droite. La sortie Y_1 est donc une fonction des entrées (X_1 ; X_2). Par exemple, Y_1 peut être la valeur de la turbidité le lendemain, X_1 le débit de la rivière Outaouais la veille, et X_2 la vitesse moyenne du vent à Dorval le jour même. La transformation identité (i.e. aucune transformation n'est effectuée) est souvent utilisée dans la couche d'entrée, les variables d'entrées ayant été préalablement prétraitées. La couche cachée un et la couche de sortie suivent les transformations décrites ci-après.

Il existe aussi des réseaux de neurones dits récurrents où la sortie d'un ou plusieurs neurones est rebouclée en tant qu'entrée dans le même neurone, et/ou ces prédécesseurs. Utilisé couramment en électronique, la rétroaction fournit une 'mémoire' au système. Ainsi, ces réseaux dynamiques ont vu une utilisation croissante dans les problèmes de séries temporelles (Maier et Dandy, 2001). Cependant, la calibration de ce type de réseau s'avère plus compliquée qu'avec un réseau statique comme le « *feedforward* ». Une manière de procurer une certaine 'mémoire' aux réseaux statiques est d'inclure en entrée des variables retardées temporellement (Govindaraju, 2000a).

Le cas particulier de zéro neurone dans la couche cachée et d'une fonction d'activation linéaire dans la couche de sortie représente un modèle linéaire de base.

L'inclusion de couches cachées additionnelles, entre la couche cachée un et la couche de sortie, peut augmenter la complexité de la sortie modélisée (Bishop, 1995). Toutefois, un PMC avec une seule couche cachée est suffisant pour approximer n'importe quelle fonction continue sous réserve de le construire avec suffisamment de

complexité, i.e. de neurones cachés [(Hornik et al., 1989) cités par (Coulibaly et al., 1999; Haykin, 1999)].

Il est important de noter que la recherche sur les réseaux de neurones reste un domaine ouvert. Effectivement, le théorème ci-dessus dit qu'une solution existe, mais ne dit pas avec quelle méthode l'atteindre ! À ce jour il n'existe pas de méthode universelle de résolution de problèmes par réseau neuronal qui fasse l'unanimité au sein de la communauté scientifique. Le choix d'une méthode est spécifique au problème considéré, certaines méthodes ayant donné de meilleurs résultats empiriques dans des applications précises.

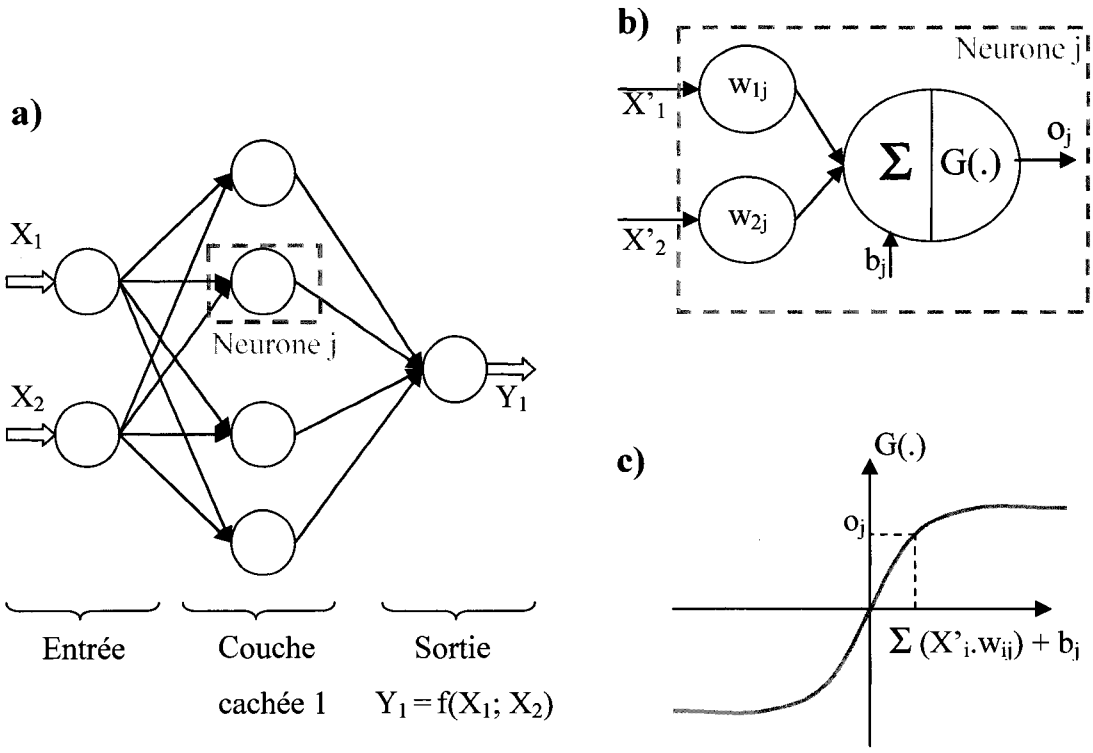

Figure 1-4 : Exemple de RNA de type perceptron multicouches "feedforward" (a) schéma d'ensemble d'un réseau 2:4:1. (b) détail du neurone j. (c) détail de la fonction d'activation G(.)

Sur la Figure 1-4 b, nous observons le détail du neurone j. En entrée de cette unité élémentaire les variables prétraitées X'_i sont pondérées par les scalaires w_{ij}, scalaires appelés poids de connexion. L'ensemble est ensuite sommé avec l'ajout d'un terme additionnel b_j appelé biais du neurone j. Cette somme pondérée plus le terme de biais est transformée par une fonction d'activation G(.). Si M représente le nombre de neurones dans la couche précédant le neurone j (ici M est le nombre d'entrées, soit deux), alors la sortie du neurone j (o_j) s'écrit donc :

$$o_j = G\left(\sum_{i=1}^{M} w_{ij} \times X_i' + b_j\right) \qquad (1\text{-}1)$$

Les paramètres libres du modèle, paramètres à étalonner lors de la phase de calibration du modèle (phase dite d'apprentissage), sont les variables $\{(w_{ij})_{i=1...M} ; b_j\}_{j=1..L}$. Où L représente le nombre total de neurones du réseau. Augmenter la complexité du modèle revient à inclure plus de neurones, et rajouter plus de paramètres libres à étalonner, il faut donc plus de données disponibles ($\underline{\mathbf{X}}$; $\mathbf{Y}$) pour le faire. Cette limitation au pouvoir de modélisation des RNA nécessite de bâtir des modèles les plus parcimonieux possible (Haykin, 1999).

Sur la Figure 1-4 c, le détail de la fonction d'activation G(.) est donné, dans le cas particulier où G est la tangente hyperbolique. Cette fonction peut être choisie de type créneau (quoique jamais utilisée), linéaire, logistique ou tangente hyperbolique. Les deux dernières procurant une transformation non linéaire en sortie des neurones permettant de modéliser un lien non linéaire entre entrées et sortie (Govindaraju, 2000a), et sont les fonctions les plus utilisées dans la couche cachée (Maier et Dandy, 2000). Notons qu'un simple changement de variable affine permet de passer de l'une à l'autre. Concernant la couche de sortie, certains auteurs ont recours à la fonction de transfert linéaire dans des applications où il est nécessaire d'extrapoler au-delà des données qui ont servi à la calibration des paramètres libres du modèle (Maier et Dandy, 2000).

Phase de calibration ou d'apprentissage

Comme mentionné ci-dessus, chaque neurone ou entrée supplémentaire augmente le nombre de paramètres libres du modèle. Il convient de trouver la combinaison de ces paramètres qui minimise l'erreur du critère de performance considéré. L'ajustement des poids des connexions et des biais est appelé phase d'apprentissage du modèle.

Il existe deux types d'apprentissage : non supervisé et supervisé. L'apprentissage non supervisé ajuste les poids du modèle en ne tenant compte que des entrées des données ($\underline{\mathbf{X}}$), indépendamment de la sortie réellement observée ($\mathbf{Y}$). Le réseau apprendra par là une représentation capturant les similitudes ou différences des données (Coulibaly et al., 1999). Ce type de réseau est principalement utilisé en reconnaissance des patrons (« *clustering* ») avec les cartes auto-organisatrices. Ces cartes permettent de réduire l'espace de n entrées (soit $\Re^n$) dans un espace de dimension inférieur (Dreyfus et al., 2004).

Ce type d'apprentissage ne sera pas utilisé, au profit de l'apprentissage dit supervisé. Cette fois-ci, les couples entrées et sortie de chaque exemple ($\mathbf{X}_j$, Y_j) sont connus et le réseau ajuste itérativement ces paramètres libres afin de minimiser la fonction d'erreur considérée. Le plus souvent il s'agit de l'erreur quadratique moyenne qui est l'espérance du carré de la différence entre la valeur réellement observée et celle prédite par le modèle (Maier et Dandy, 2000). Ce type d'apprentissage vise à extraire une approximation numérique liant entrées et sortie dans le cas d'un modèle de régression. Dans le cas d'un modèle de classification, l'apprentissage supervisé permet d'établir l'équation de la ou des frontières séparant une ou plusieurs classes.

L'algorithme d'apprentissage le plus répandu est l'algorithme de rétropropagation (ARP) (Coulibaly et al., 1999; Maier et Dandy, 2000). Le principal objectif de l'ARP est de modifier les paramètres libres du réseau afin de minimiser l'écart entre valeurs prédites et valeurs observées. Un défi partiellement résolu par l'ARP et celui

d'allocation spatiale : connaissant l'erreur commise sur la prédiction, quels neurones faut-il aller modifier et de combien ?

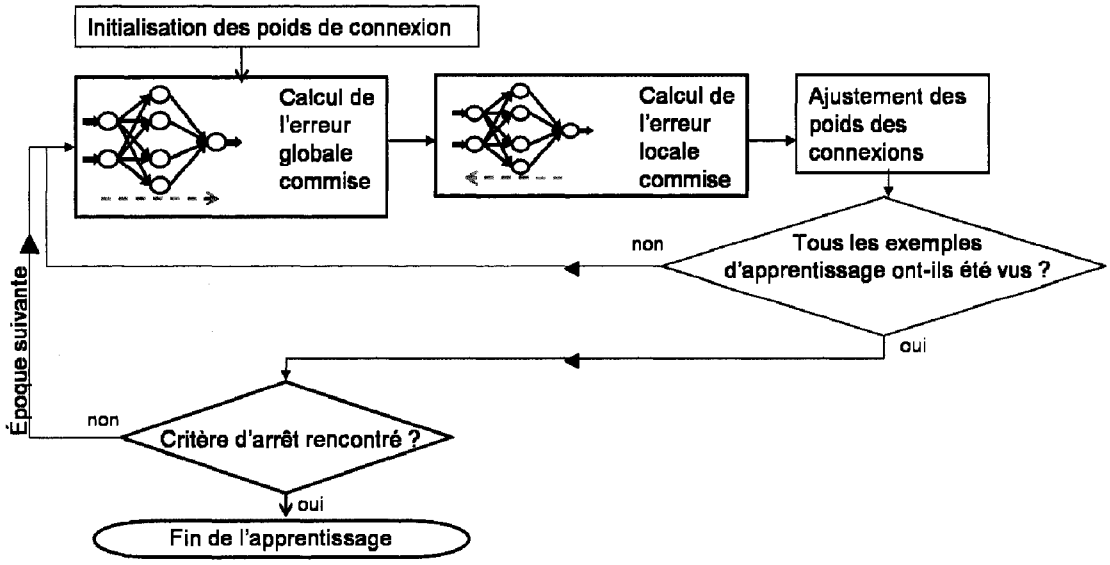

Figure 1-5 : Schéma simplifié de l'apprentissage du RNA par rétropropagation

Voici une description sommaire du fonctionnement séquentiel de l'ARP[2] (Figure 1-5). Initialement, tous les paramètres libres sont initialisés suivant une distribution (uniforme ou gaussienne) de moyenne zéro et de variance un, constituant une initialisation aléatoire des paramètres libres qui revêt une grande importance dans la solution finale retenue, nous en discuterons plus bas. Les poids des connexions ont des valeurs $(\mathbf{w})^0$ à l'étape initiale. Dans un premier temps l'ARP fonctionne dans le sens aval. Les exemples réservés pour l'apprentissage sont présentés au réseau, et les valeurs de sortie pour chaque exemple sont comparées à la valeur cible (valeur de sortie réellement observée), ceci donne la moyenne de l'erreur commise. Ensuite, l'ARP fonctionne en sens inverse (sens amont), et propage l'erreur de la sortie vers les entrées. Pour chaque neurone, en commençant par celui de la couche de sortie

[2] Plus de détails sont disponibles dans (Haykin, 1999)

jusqu'aux entrées, l'algorithme calcule la dérivée partielle de l'erreur moyenne commise par rapport aux paramètres libres du neurone en question. Dans l'espace des paramètres libres du réseau, l'algorithme calcule ainsi la surface d'erreur, localement au voisinage du point $(\mathbf{w})^0$. Puis, vient la troisième étape, la phase d'ajustement des paramètres libres du réseau. À partir du gradient de la surface d'erreur au voisinage du point $(\mathbf{w})^0$, l'ARP va chercher la direction de plus grande pente afin de converger vers un minimum d'erreur. Notons que les améliorations potentielles à l'ARP vont être sur la manière de converger vers un minimum de la surface d'erreur (suivant la ligne de plus grande pente par exemple). L'avancée dans la direction de plus grande pente est proportionnelle au gradient de l'erreur commise en $(\mathbf{w})^0$, soit $(\nabla E_{|(\mathbf{w})}{}^0)$. Le coefficient de proportionnalité s'appelle le taux d'apprentissage (η). À la fin de la première itération, le point sur la surface d'erreur se déplace en $(\mathbf{w})^1$ tel que :

$$(\mathbf{w})^1 = (\mathbf{w})^0 - \eta \,.\nabla E_{|(\mathbf{w})}{}^0 \qquad (1\text{-}2)$$

Tous les exemples réservés pour l'apprentissage sont ainsi présentés itérativement. Ceci est appelé une époque de l'algorithme d'apprentissage. Plusieurs époques peuvent être présentées au réseau jusqu'à ce que l'algorithme rencontre un des critères d'arrêt de l'apprentissage.

La forme de la surface d'erreur étant spécifique à chaque problème, un taux d'apprentissage optimal doit être trouvé pour chaque application. En effet, si η est trop petit, une faible distance sera parcourue à chaque étape, la convergence sera longue (surtout sur un plateau de la surface d'erreur où les pentes sont quasiment nulles), et l'algorithme risque fort de se retrouver coincé dans un minimum local de la surface d'erreur. Si η est trop élevé, de grandes enjambées seront parcourues sur la surface d'erreur, au risque de zigzaguer par-dessus certains minima (crevasses).

Bien que simple en termes de calcul et demandant peu de mémoire, il semble évident que, selon la forme de la surface d'erreur, l'ARP peut être très lent à converger. Ainsi,

plusieurs améliorations à l'algorithme présenté ci-dessus sont disponibles. Citons notamment l'apprentissage adaptatif qui fait varier le taux d'apprentissage (η) en fonction de la direction de déplacement (η est augmenté si l'algorithme se déplace plusieurs fois dans le même sens, et réduit s'il zigzague). L'apprentissage adaptatif permet de trouver un η quasi optimal. Afin d'accélérer la convergence, l'inclusion d'un terme du deuxième ordre (dérivée seconde de l'erreur) permet de tenir compte de la courbure de la surface d'erreur. À l'Équation (1-2) est rajouté un terme de momentum (μ) tel que :

$$(\mathbf{w})^{n+1} = (\mathbf{w})^{n} - \eta .(\nabla E)_{(\mathbf{w})^{n}} + \mu . \left[\frac{\partial^2 E}{\partial \mathbf{w}^2}\right]_{(\mathbf{w})^{n}} \qquad (1\text{-}3)$$

D'autres algorithmes plus complexes existent. Ils sont toutefois plus lourds en termes de calcul et de mémoire requise, mais plus performants et plus rapides pour la recherche de minima. Citons notamment les méthodes du deuxième ordre de Levenberg-Marquardt, ou les algorithmes ne se déplaçant pas suivant la ligne de plus grande pente (méthodes du gradient conjugué). Empiriquement, on observe que la première méthode convient mieux aux problèmes de régression avec la fonction d'erreur quadratique moyenne, alors que les méthodes du gradient conjugué donnent de meilleurs résultats de convergence pour les problèmes de classification (The MathWorks, 2007).

Un des principaux désavantages des réseaux de type PMC est que la convergence vers un minimum local dépend de l'initialisation des poids au début de l'algorithme $(\mathbf{w})^{0}$: i.e., selon le point de départ sur la surface d'erreur, l'algorithme va converger vers différentes solutions. Ainsi, le même réseau entraîné deux fois de suite ne donnera pas les mêmes résultats. On dit que les résultats obtenus sont probabilistes.

1.3.2 Revue des applications en hydrologie

Après avoir rappelé brièvement les concepts fondamentaux des réseaux de neurones de type perceptron multicouches, voici un aperçu de l'utilisation qui en a été faite dans le domaine de l'hydrologie, et des éléments de méthodologie préconisés par les auteurs. Tout d'abord une section traitera des articles faisant une revue des paramètres utilisés et mettant en avant les principales phases nécessaires au développement d'un modèle par réseau neuronal. Puis, dans une deuxième partie, les éléments de méthodologie suggérés par divers auteurs seront abordés points par points.

Revues et état de l'art

La revue de littérature fait ressortir quatre articles faisant le point sur l'utilisation de réseaux de neurones en modélisation hydrologique et environnementale.

Le premier article fait la synthèse de deux applications concluantes de l'auteur en modélisation de la qualité microbiologique de l'eau (Brion et Lingireddy, 2003). Un modèle RNA servit à prédire les pointes de concentration de *Cryptosporidium* et de *Giardia* à la prise d'eau brute de la rivière Delaware, USA (Neelakantan et al., 2001) avec un taux de classification correcte sur l'ensemble de Test de 88% et 94% respectivement. Ces travaux permettraient de prédire des données microbiologiques difficiles à mesurer à partir d'indicateurs plus aisément mesurables (*E. Coli*, coliformes fécaux, paramètres météorologiques, débit de la rivière, turbidité, etc.). Le deuxième modèle vise à discriminer parmi des sources et des âges de pollutions fécales : égout humain (frais) ou ruissellement d'eau sous influence d'animaux (pollution plus âgée). Une série de modèles discriminant entre deux classes furent assemblés en cascade pour produire la prédiction finale. L'étude des variables influentes de ce modèle permit de découvrir d'autres variables indicatrices à substituer pour la mesure de la qualité de l'eau.

Les quatre articles suivants font la revue de l'utilisation des RNA en hydrologie (Coulibaly et al., 1999; Govindaraju, 2000a; Govindaraju, 2000b; Maier et Dandy, 2000). Les domaines d'application sont variés, ils concernent : la classification des données hydrologiques, la prévision des débits, de la qualité de l'eau, de la consommation en eau, des apports annuels en eau et en sédiments aux réservoirs, et la production hydroélectrique (Coulibaly et al., 1999).

Par ailleurs, l'utilisation des RNA au sein des usines de filtration d'eau potable a pour but d'améliorer la qualité de l'eau produite tout en diminuant les coûts de production, les RNA étant implantés au sein de systèmes automatisés de contrôle des procédés (Baxter et al., 2001). L'étude de cas présentée est basée sur les deux usines de filtration d'Edmonton (AB). Voici des exemples de l'utilisation des modèles RNA : prédiction de la couleur à l'eau brute, estimation de la demande en eau, prédiction de l'enlèvement de couleur et de turbidité après filtration assistée par coagulant et charbon actif, estimation de la dose d'alun requise (pour atteindre un niveau de turbidité donné à l'effluent), estimation de la dureté totale à l'effluent et estimation de la dose de chaux requise, et prédiction de l'enlèvement de particules au travers des filtres (autre indicateur de la performance de filtration). Tous ces modèles RNA furent implantés dans un simulateur virtuel destiné à former les opérateurs sur la station. De plus, notons que l'American Water Works Association vient de publier sur ce sujet un guide intitulé « *Real Time Artificial Intelligence Control and Optimization of a Full Scale WTP* » (2007).

Phases à considérer lors de l'élaboration d'un modèle RNA

Tous les auteurs concluent qu'aucune méthodologie universelle n'existe, mais fournissent un cadre général d'étapes à suivre pour l'élaboration d'un modèle RNA. Ce cadre est aussi décrit dans deux autres références (Dreyfus et al., 2004; Maier et

Dandy, 2001). Voici les grandes phases à considérer, elles sont regroupées en onze étapes.

1. Définition des objectifs

Avant d'envisager toute modélisation, il semble essentiel de bien détailler les objectifs auquel devra répondre le modèle : quelle doit être la donnée en sortie? Mesurée avec quel pas de temps ? Y a-t-il un niveau de performance acceptable minimal à atteindre ? Quel critère de performance faut-il choisir ? Est-ce la précision de la prédiction, la rapidité d'apprentissage, ou bien la précision temporelle de la prévision, i.e. aucun retard n'est admis (Maier et Dandy, 2000).

2. Quel type de modèle (linéaire, RNA, etc.) est le mieux adapté ?

En fonction des objectifs précédents, il est possible de choisir quel modèle parmi ceux disponibles est le mieux adapté aux besoins de l'utilisateur. La modélisation de phénomènes non linéaires recommande l'usage de RNA. Selon le critère de performance adopté et la variable à modéliser, diverses architectures RNA pourront être employées, chacune ayant ses avantages et inconvénients.

3. Récupération de la base de données

La constitution d'une base de données préliminaire se fait par connaissance préalable du sujet, et/ou disponibilité des données (Tremblay, 2004). Toute connaissance préalable sur le système à modéliser oriente la sélection des variables vers un premier groupement de variables candidates pour le modèle (Bowden et al., 2005).

4. Tri de la base de données

Cette partie découle de la précédente. Les données récupérées provenant de divers formats, il convient de tout uniformiser dans un seul fichier. Les exemples sur lesquels

le modèle RNA va se calquer doivent être représentatifs de la population d'exemples en général. L'inclusion de patrons non représentatifs pourrait perturber le réseau dans son apprentissage. Ainsi, certains auteurs éliminent les « *outliers* » de leur base de données (Govindaraju, 2000a). Notons que, dans le cas présent, nous cherchons à modéliser les évènements extrêmes, les valeurs 'aberrantes' seront conservées afin de fournir le plus d'exemples possibles nécessaires à leur bonne prédiction.

5. Choix des variables d'entrées

Le choix des variables d'entrées se décompose en deux parties : la réduction de la dimension de l'espace des entrées, et l'élimination des entrées non pertinentes.

Ces deux étapes sont cruciales, car l'inclusion d'un nombre insuffisant de variables ne permet pas d'atteindre une performance maximale (les variables non inclues peuvent contenir de l'information utile à la prédiction). En revanche, inclure trop de variables dégrade aussi la performance : l'inclusion de variables ne procurant aucune information supplémentaire brouille le réseau dans sa recherche de la loi liant entrées et sortie du modèle. Ceci augmente la complexité du réseau et dégrade sa capacité de généralisation : plus d'exemples sont gaspillés pour éliminer les variables non pertinentes.

Réduction de la dimension de l'espace des entrées

Afin de tendre vers l'approximation parcimonieuse, il est nécessaire de regrouper les entrées véhiculant une information identique, ceci permet d'éliminer la redondance de l'information. Les techniques les plus couramment utilisées sont les analyses en composantes principales (ACP), ou plus rarement les cartes auto-organisatrices. Les ACP consistent à projeter les données dans des plans principaux expliquant le plus de variance; ces plans sont construits orthogonaux, si bien que les composantes projetées

sont non corrélées entre elles. Toutefois, de l'information peut être perdue lors de la projection (Dreyfus et al., 2004).

Elimination des variables d'entrée non pertinentes

Une première méthode proposée est celle dite du descripteur sonde (Dreyfus et al., 2004). Pour choisir les entrées d'un modèle neuronal, on considère dans un premier temps les variables pertinentes pour un modèle linéaire (régression polynomiale ou analyse discriminante par exemple), puis on utilise les variables sélectionnées dans le RNA. Le modèle linéaire est calibré avec toutes les entrées potentielles plus une variable générée aléatoirement n'ayant aucun lien avec la sortie (le descripteur). Une orthogonalisation permet de trier les variables par ordre de pertinence sur la base de la corrélation avec la sortie. Toutes les variables moins bien classées que le descripteur sont éliminées. En répétant ces étapes avec un nombre différent de variables descriptives, on se définit un niveau de risque acceptable de retenir une entrée non pertinente. Le recours à un modèle linéaire garantit l'unicité de la solution trouvée, est léger en termes de calculs, et permet de traiter un grand nombre d'entrées en même temps. Cette méthode permet un classement des entrées et théorise la sélection des entrées sur la base du risque d'acceptation d'une entrée non pertinente.

De plus, deux articles traitent de la sélection des entrées (Bowden et al., 2005; Maier et Dandy, 1997). Le premier recense dans les articles étudiés cinq méthodes pour choisir les entrées pertinentes et en propose deux autres. Ces cinq méthodes sont : l'utilisation de connaissance préalable sur le système, les méthodes basées sur la corrélation croisée (composante linéaire liant des variables), les méthodes heuristiques (essais et erreurs), l'utilisation des RNA par analyse de sensibilité, et la composition des cinq méthodes précédentes. Les méthodes précédentes (incluant le descripteur sonde) présentent souvent l'inconvénient de ne considérer que la composante linéaire liant entrées et sortie, alors que le phénomène à modéliser est non linéaire.

Les deux autres méthodes proposées par l'article sont particulières. La première est libre de modèle et utilise la notion statistique de « *partial mutual information* » pour estimer les interdépendances, linéaires ou non, entre entrées et sortie. La méthode permet de calculer pour chaque entrée la réduction d'incertitude sur la variable de sortie y sachant la variable d'entrée **X**. Un classement des entrées peut ensuite être mené. La deuxième méthode suggère l'utilisation des performances de RNA entraînés pour déterminer quelles entrées améliorent le mieux la prédiction. Cette méthode est issue de la composition de trois algorithmes d'apprentissage automatisés : les cartes auto-organisatrices (SOM), les algorithmes génétiques (GA) et les réseaux de neurones de régression généralisée (GRNN). Pour chaque combinaison d'entrées, les SOM permettent de réduire la dimension des entrées, puis un réseau de type GRNN est entraîné afin d'évaluer la performance du groupe d'entrées en question sur la sortie. L'utilisation d'un GRNN au lieu d'un PMC est justifiée par l'unicité de la solution obtenue avec le GRNN au prix d'un temps d'apprentissage plus long qu'avec le PMC. Si peu d'entrées potentielles doivent être testées, il est aisé de tester de manière exhaustive toutes les combinaisons d'entrées possibles; malheureusement, ce n'est pas toujours le cas. Ainsi, les GA permettent de retenir une combinaison d'entrées potentielles optimale sans avoir à tester toutes les combinaisons possibles. Une partie de cette méthode sera utilisée par la suite en tant que deuxième avis pour la sélection des entrées du modèle.

6. Partitionnement des exemples

Note : si les RNA sont utilisés pour le choix des entrées, alors cette étape doit être menée avant.

Il est d'usage courant en modélisation de séparer les données en deux ensembles : l'apprentissage et le test. Le premier ensemble, apprentissage, sert à étalonner les

paramètres libres du réseau; le deuxième ensemble, test, sert à évaluer la capacité de généralisation du modèle avec des données qui n'ont jamais été présentées au modèle.

En modélisation RNA, l'ensemble d'apprentissage est encore découpé en deux : apprentissage, et sélection. Comme précédemment, le sous-ensemble d'apprentissage sert à étalonner les paramètres libres du réseau. Alors que le sous-ensemble de sélection va être employé afin de maximiser un paramètre de l'algorithme d'apprentissage du réseau : le nombre d'époques d'apprentissage, donc l'arrêt de ce dernier. Cette technique permet d'éviter le phénomène de sur-apprentissage (le réseau, à force de voir les exemples d'apprentissage, s'ajuste sur le bruit des données et perd en généralisation).

Le fonctionnement de la méthode de validation croisée est schématisé comme suit (Figure 1-6) : au fur et à mesure que les exemples de l'ensemble d'apprentissage sont présentés au réseau, celui-ci adapte ces poids pour minimiser à chaque étape l'erreur commise. Cette courbe est ainsi décroissante (trait plein). L'observation de l'erreur sur l'ensemble de sélection montre deux phases (trait interrompu). La première où le réseau est en train de capturer le phénomène physique sous-jacent montre que l'erreur diminue. Puis, la deuxième montre que l'erreur augmente car le réseau copie uniquement l'ensemble d'apprentissage : on parle de sur-ajustement. L'erreur de sélection passe par un minimum. La valeur des poids de connexion étant conservée à chaque étape, l'algorithme d'apprentissage par la méthode de validation croisée retiendra finalement la configuration à ce minimum. Il s'agit d'un compromis entre un bon apprentissage du phénomène physique à modéliser (faible erreur d'apprentissage) et une bonne généralisation (minimisation de l'erreur de sélection).

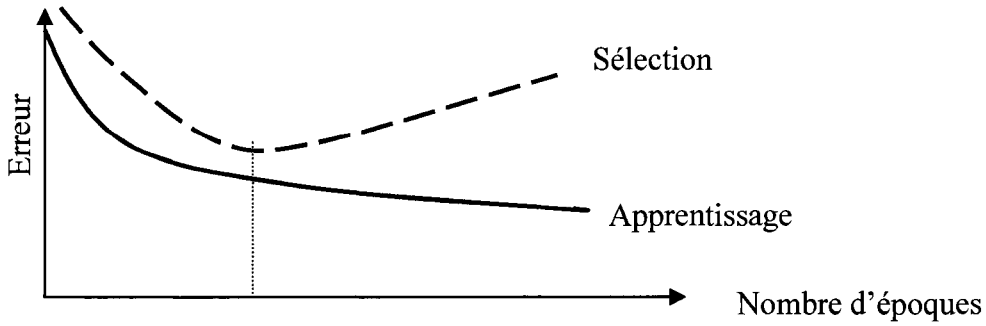

Figure 1-6 : Erreur d'apprentissage et de sélection en fonction du nombre d'époques[3]

Dans la plupart des articles recensés dans la revue sur l'utilisation des modèles RNA en hydrologie, l'ensemble d'apprentissage comprend de 50 à 80% des données totales disponibles ; sélection et test n'ayant que de 10 à 25% des données restantes (Maier et Dandy, 2000).

Les RNA performent mieux s'ils ne sont pas nourris avec des données au-delà de celles utilisées pour la calibration (Shahin et al., 2004). L'ensemble d'apprentissage doit évidemment contenir des données représentatives du phénomène à modéliser, mais en plus il doit si possible contenir les exemples extrêmes de la base de données. Cependant, les ensembles de sélection et de test doivent aussi représenter la même population : comment évaluer la performance de généralisation d'un modèle prédictif si les données test utilisées sont différentes du phénomène à modéliser ?

Trois méthodes générales sont proposés pour l'échantillonnage (Shahin et al., 2004). La première, la plus répandue, est aléatoire et ne tient pas compte des statistiques des données. Une proportion fixée de données est répartie aléatoirement entre les trois ensembles. Bien souvent le découpage est plus rudimentaire que cela : pour des séries temporelles, c'est le temps qui est découpé en trois parties, l'ensemble de test étant souvent la période la plus récente. Une deuxième méthode prend toujours ces

[3] Inspiré de Haykin (1999)

exemples de manière aléatoire dans des proportions fixées, mais vérifie a posteriori que les trois ensembles soient représentatifs d'une même population (qu'il n'y ait pas de différences statistiquement significatives entre les ensembles). Le modélisateur procède par essais et erreurs jusqu'à obtenir un bon échantillonnage. Enfin, la troisième méthode regroupe les données par leurs similitudes (« *clustering* ») avant de les répartir dans les ensembles. Ceci assure que la répartition obtenue donne lieu à trois ensembles représentatifs d'une même population. Cette dernière méthode semble montrer les meilleurs résultats (Shahin et al., 2004).

Afin de s'assurer que les performances obtenues pour un réseau soient indépendantes du partitionnement des exemples utilisés, il est recommandé d'avoir recours à plusieurs échantillonnages. Par exemple, cinquante réseaux entraînés avec répartitions différentes ont été utilisés afin d'évaluer la déviation obtenue sur la valeur de sortie prédite (10^e et 90^e centile) pour chaque modèle (Valentin et al., 1999).

7. Choix d'une architecture de réseau

L'architecture va définir le nombre de paramètres libres du réseau. Il s'agit de la manière dont l'information s'écoule à travers le réseau (réseau récurrent ou « *feedforward* »), si le réseau est totalement ou partiellement connecté, et de sa géométrie. On appelle géométrie le nombre de couches cachées et le nombre de neurones dans chaque couche cachée; les couches d'entrées et de sortie ayant déjà été fixées auparavant.

Il s'agit d'une phase tout aussi importante que la sélection des entrées pertinentes car elle fixe la complexité du modèle élaboré. La plupart des articles recensés utilisent des réseaux de type PMC « *feedforward* » totalement connectés. Dans la majeure partie des cas, la géométrie est déterminée par essais et erreurs (Maier et Dandy, 2000).

Récemment quelques articles essaient d'adopter des méthodes plus systématiques pour la détermination de la géométrie, mais cette approche n'est pas universellement effectuée. Une des méthodes est l'élagage des connexions (« *pruning* » en anglais). Cette méthode considère au départ un réseau plus grand que nécessaire, et va éliminer les connexions dont la contribution est la plus faible. Si le poids de la connexion tombe en dessous d'un certain seuil prédéterminé, la connexion est éliminée. Le réseau final obtenu est partiellement connecté. Ces méthodes aboutiraient à un réseau plus parcimonieux, donc ayant une meilleure capacité de généralisation (Corani et Guariso, 2005).

De plus, ont été inclus dans le terme architecture les transformations pré et post traitement effectuées sur les entrées et les sorties des modèles, ainsi que les fonctions d'activation de chaque neurone. Plus de détails sur les objectifs et la manière de transformer les variables sont donnés dans l'Annexe D.

8. Calibration des réseaux (phase d'apprentissage)

Le réseau 'apprend' de ses expériences passées. La phase d'apprentissage a donc pour but d'étalonner les paramètres libres d'un modèle RNA dans le but de minimiser une fonction d'erreur, fonction calculée sur la base des exemples disponibles. Une description sommaire de la phase d'apprentissage a déjà été donnée dans la section 1.3.1).

Il convient dans cette phase de choisir un algorithme d'apprentissage (rétropropagation, Levenberg-Marquardt, gradient conjugué, etc.) et pour chaque algorithme de déterminer les constantes d'apprentissage. Par exemple, pour l'algorithme de rétropropagation, ces constantes sont le taux d'apprentissage (η), le momentum (μ), et le nombre maximal d'époques à considérer. L'effet des paramètres internes de l'algorithme de rétropropagation est très bien documenté dans une étude empirique sur la prédiction de la salinité de la rivière Murray en Australie (Maier et

Dandy, 1998a; Maier et Dandy, 1998b). La performance des méthodes d'apprentissage est comparée empiriquement pour la même étude de cas (Maier et Dandy, 1999). Il est aussi possible de consulter l'étude de cas fournie dans l'aide du logiciel Matlab® (The MathWorks, 2007). Il ressort empiriquement que les algorithmes de type gradient conjugué performent mieux pour les problèmes de classification; alors que, pour la régression l'algorithme de Levenberg-Marquardt est préférable.

Les algorithmes d'apprentissage ont pour but de minimiser une fonction de coût. Les fonctions de coût les plus utilisées sont l'erreur quadratique moyenne pour la régression, et l'entropie croisée pour les problèmes de classification (Statsoft, 2006).

Il faut toutefois faire attention, la recherche d'un minimum de la fonction de coût peut converger vers des solutions physiquement improbables (par exemple, la turbidité à Des Baillets inversement proportionnelle à la turbidité en amont à Beauharnois). Ainsi des termes de contraintes peuvent être ajoutés à la fonction de coût pour pénaliser des contributions improbables de certaines entrées si une connaissance préalable du système à modéliser est disponible (Kingston et al., 2005).

9. Choix d'un critère de performance

Une fois une série de modèles étalonnés, il est nécessaire de les comparer entre eux afin de favoriser celui répondant le mieux aux besoins de l'utilisateur. Le choix d'un critère de performance adéquat permet de quantifier la performance de chaque modèle et des les trier par ordre d'efficacité.

Le critère de performance comprend un terme principal, et au besoin des termes additionnels de moindre importance. La contribution de chaque terme est pondérée par l'importance relative qu'à ce terme vis-à-vis des objectifs du modèle, dépendamment de l'appréciation du modélisateur. Plusieurs critères principaux sont utilisés couramment en hydrologie. Il y a par exemple pour les problèmes de régression,

l'erreur des moindres carrés sommés, l'erreur absolue moyenne, ou le coefficient de Nash-Sutcliff (Martinec et Rango, 1989). L'adéquation temporelle entre les prédictions et les observations peut être calculée par le coefficient de corrélation de Pearson (r) (Tremblay, 2004). Pour la discrimination, le pourcentage de classification correcte semble être le critère le plus intuitif, bien que des matrices de perte peuvent pénaliser plus fortement certaines erreur de classification (Statsoft, 2006).

En plus de ces critères de performance principaux, des termes additionnels de plus faible pondération peuvent être ajoutés. Il peut s'agir de termes pénalisant les modèles trop complexes, ou bien pénalisant l'inclusion d'entrées difficilement disponibles (Tremblay, 2004).

10. Détermination du meilleur réseau

Les critères de performance permettent d'ordonner les réseaux candidats, il convient de retenir le ou les modèles répondant le mieux aux besoins du problème.

11. Bâtir le modèle final

Cette partie vise à fixer les détails techniques pour l'implantation du modèle en station, à savoir : le code informatique, l'interface graphique pour les utilisateurs, la récupération des bases de données, et au besoin, la mise en commun des prédictions des modèles si plusieurs modèles ont été retenus.

1.3.3 Quelques exemples spécifiques

Au cours de cette section, une série d'articles présentant des similitudes avec ce projet seront commentés. La section est divisée en deux parties : la première partie traite de la prédiction de la qualité à l'eau brute; puis la seconde partie porte sur la prédiction de la qualité de l'eau filtrée, bien souvent avec ajout de coagulant dans la filière de filtration.

Prédiction de la qualité de l'eau brute

Une série d'articles furent déjà commentés par Tremblay (2004), par souci de diversification, d'autres articles seront étudiés ici.

Prédiction de débit de rivière

Un PMC de type « *feedforward* » fut utilisé pour prédire récursivement de un à sept jours à l'avance le débit de la rivière Mistassibi (QC) (Birikundavyi et al., 2002). Trente deux années de données journalières furent récoltées et découpées par années en trois ensembles de 17, 12 et 3 ans. Les entrées retenues furent les mêmes que le modèle linéaire ARMAX afin de pouvoir comparer les performances des deux modèles; à savoir : les débits des jours précédents, la température moyenne prédite pour le lendemain, la fonte des neiges calculée, et la précipitation. Les 13 entrées du modèle furent normalisées selon la méthode du min max. L'algorithme d'apprentissage utilisa la rétropropagation avec arrêt lorsque l'EQM est minimale et le coefficient de Nash-Sutcliff maximal sur l'ensemble de sélection. Sur les années de test ces coefficients montrent des performances allant respectivement de 17,71m^3/d à 95,89 m^3/d, et de 0,9930 à 0,7969. La plus mauvaise performance étant obtenue pour les modèles prédictifs à 7 jours : la prédiction étant déterminée à partir des prévisions les jours précédents, l'erreur s'accumule donc de jours en jours.

Une série de commentaires intéressants viennent compléter cet article (Sudheer et al., 2004). Cigizoglu suggère le recours aux réseaux GRNN pour éviter les minima locaux lors de l'apprentissage. Les sorties de ces réseaux étant bornées par les valeurs extrêmes de l'ensemble d'apprentissage, nous obtenons toujours des résultats physiquement plausibles. Cependant, ces réseaux extrapoleraient moins bien que les PMC. Le problème de minimum locaux peut être réduit en considérant un taux d'apprentissage et un momentum propre à chaque neurone. Une des améliorations suggérées est aussi la construction de sept modèles prédictifs pour la prédiction de un

à sept jours. Ceci éviterait d'accumuler l'erreur au fur et à mesure des prédictions. De plus, l'inclusion d'une composante périodique en entrée supplémentaire guiderait le réseau vers la reconnaissance de phénomènes saisonniers (fonte des neiges, pluies d'automne, etc.). La dernière amélioration proposée est basée sur la méthode du « *range dependent neural network* ». Cette méthode vise à découper les données en intervalles (bas, moyen, élevé) et d'attribuer des règles d'apprentissage différentes selon la classe considérée. La performance de prédiction des évènements élevés étant limitée par leur faible occurrence (soit peu d'exemples élevés disponibles), il convient de leur accorder plus de poids à l'apprentissage.

Prédiction des concentrations en sédiments

Deux autres auteurs ont traité la prédiction de concentration en sédiments.

Dans un premier temps, deux articles visent à prédire la qualité de l'eau dans un bassin versant de la forêt boréale canadienne, en Alberta. Il s'agit de la prédiction des concentrations en phosphore et débits dans des zones où aucune jauge de mesure n'est présente (Nour et al., 2006a), mais aussi du débit de la rivière et de la concentration de matières en suspension (MES) (Nour et al., 2006b). Les entrées des modèles sont divisées en deux catégories : entrées ayant un lien de cause à effet avec la sortie, entrées décalées temporellement, et des entrées reflétant un cycle annuel ou saisonnier. Une analyse spectrale des variations des paramètres permet d'identifier que les variations mensuelles représentent la période dominante, d'où la génération de signal d'entrée de forme sinusoïdale et de même période. Une analyse par corrélation croisée identifie les décalages temporels à considérer. Les autres entrées potentielles pour le débit sont les précipitations, les composantes périodiques, les températures de l'air, l'indice de température degré jour, et l'estimation de la fonte des neiges. Les concentrations en phosphore incluent les prévisions sur le débit, les concentrations en phosphore des jours précédents, la température de l'air et les composantes périodiques.

De même que ci-dessus, l'apprentissage a lieu avec l'algorithme de rétropropagation avec méthode de validation croisée, et la performance des modèles est déterminée par le coefficient de détermination (R^2) ainsi que l'EQM. La division des données est opérée par « *clustering* » en répartissant les données dans un ratio 3 :1 pour l'apprentissage et l'ensemble de test. L'architecture et les paramètres d'apprentissage optimaux furent déterminés par essais et erreurs. Les résultats sur la prédiction des MES atteignent un R^2 de 0,91 sur l'ensemble de test. Le modèle sous-estime légèrement les pointes lors de la fonte des neiges.

Le deuxième auteur mérite une attention particulière car il travaille depuis plus de dix ans sur le développement de modèles prédictifs des concentrations en sédiments, citons notamment les publications suivantes.

Les premiers articles utilisent des réseaux PMC avec algorithme de rétropropagation. Les entrées du réseau, les paramètres d'apprentissage et l'architecture du réseau sont tous déterminés par essais et erreurs. Les entrées sont normalisées entre 0,2 et 0,8 par la méthode min max. Des travaux antérieurs montrent, pour la rivière Tees (Angleterre), une corrélation linéaire entre turbidité et concentration en sédiments (Cigizoglu, 2002b). Les RNA dépassent la performance obtenue par les modèles conventionnels de « *sediments rating curves* » (SRC). Ces méthodes conventionnelles cherchent une solution de la forme :

$$C = a.Q^b + \varepsilon \qquad (1\text{-}4)$$

où C représente la concentration en sédiments, Q est le débit de rivière, et a et b sont des constantes à déterminer, ε est un terme d'erreur.

Cependant, il existe une hystérésis sur le lien entre sédiments et débit (Cigizoglu, 2002a), hystérésis non modélisable par l'équation ci-dessus. Ainsi, l'auteur s'est tourné vers les modèles neuronaux. Différents modèles RNA furent développés selon

la disponibilité des données. Ces modèles considèrent comme entrées les débits en amont et, si possible, les concentrations en sédiments en amont (ou les jours précédents). Les modèles RNA étalonnés sont utilisés pour estimer des concentrations en sédiments des rivières voisines afin de limiter le coût en instruments de mesure. Cette étude ne prit pas en compte les précipitations car les données ne furent pas disponibles. L'auteur suppose cependant que ce paramètre aurait pu améliorer les performances de prédiction (Cigizoglu, 2002a).

Le troisième article développe les idées précédentes pour la rivière Schuylkill à Philadelphie (Cigizoglu, 2004). La même méthodologie est utilisée : PMC, architecture déterminée par essais et erreurs, performance évaluée par R^2 et EQM... Les entrées sont déterminées par analyse statistique (corrélation croisée), et les algorithmes de rétropropagation et de Levenberg-Marquardt sont utilisés pour l'apprentissage. Cette publication vise deux objectifs : prédire la concentration en sédiments à partir des concentrations passées et amonts, et estimer la concentration en sédiments en fonction des débits passés et en amont (comme le feraient les SRC). La fonction de transfert adoptée est la tangente hyperbolique, certains modèles donnent donc des prédictions négatives à la place des faibles valeurs de concentration observées. Les modèles ayant comme entrées les concentrations en sédiments en amont performent mieux, l'auteur l'explique en disant que l'auto-corrélation des entrées décalées temporellement est plus faible que le coefficient de corrélation entre entrées et sorties.

Pour le quatrième article, il s'agit de prédire le débit journalier moyen de la rivière Ergene (Turquie) en fonction des débits passés, ou des débits de stations environnantes (Cigizoglu, 2005). Afin de résoudre les problèmes de minima locaux et de prédiction non plausibles physiquement, l'auteur utilise dorénavant un réseau de type GRNN. Il compare les résultats obtenus avec le modèle PMC le plus performant obtenu. Le choix des paramètres internes des réseaux est déterminé par essais et erreurs. Seule la

sélection des entrées mêle analyse statistique des corrélations et résultats des performances d'un modèle linéaire de régression multi variables. Les modèles linéaires, calibrés avec leurs groupes d'entrées candidates, sont classés à l'aide du critère d'Akaike. Ce critère se base sur l'erreur quadratique moyenne et inclut un terme pénalisant les modèles les plus complexes. De plus, afin d'améliorer la performance des modèles de séries temporelles, l'auteur a inclus une entrée supplémentaire sous forme de composante périodique (date en journée julienne divisée par 365).

Le cinquième article vise à prédire la concentration journalière en sédiments de la rivière Juniata (PA) en fonction des concentrations passées et des débits d'eau (Cigizoglu et Alp, 2006). Les entrées sont cette fois-ci déterminées par essais et erreurs. On retient le modèle donnant de meilleures performances sur l'ensemble de test. Les critères de performances sont toujours l'EQM et le R^2, auxquels s'ajoute la somme annuelle des sédiments. En effet, la gestion d'une retenue d'eau (par exemple dragage du fond) nécessite de connaître les apports totaux annuels. Les performances des modèles GRNN, PMC, courbes SRC, et régression linéaire multi variables sont comparées. Les résultats montrent que pour les prédictions de moyennes et fortes concentrations en sédiments, PMC et GRNN se valent (avec les critères de performances adoptés). Pour les faibles valeurs de concentrations en sédiments, les GRNN ne donnent pas de prédictions négatives.

Le sixième et dernier article recensé ici fait la synthèse des connaissances obtenues des études précédentes, et propose de nouvelles méthodes pour l'étude de cas de la prédiction de la concentration en sédiments de la rivière Schuylkill (USA) (Cigizoglu et Kisi, 2006). Le réseau utilisé est un PMC de type « *feedforward* », dont l'architecture est déterminée par essais et erreurs. L'apprentissage se fait par rétropropagation et algorithme de Levenberg-Marquardt. Les entrées et sorties sont normalisées entre 0,1 et 0,9. Les faits saillants et méthodes nouvelles proposées sont :

- L'analyse statistique (corrélation linéaire) des données pour aider au choix des entrées du modèle.
- Le « *k-fold partitionning* ». Après avoir découpé les données en k sous-ensembles, celui apportant le plus d'information est retenu. Cette méthode vise à tirer parti du fait qu'un sous-ensemble de données peut contenir plus d'informations pertinentes que toutes les données cumulées.
- Le « *range dependent neural network* ». La plage de sortie de la variable à mesurer est découpée en plusieurs classes (par exemple, bas moyen et élevé). Un modèle neuronal est développé par classes. L'idée sous-jacente est qu'il existe une dynamique différente qui gouverne les évènements bas, moyens, et hauts; d'où la construction d'un modèle par classes.

Prédiction des performances de la filtration avec coagulant

Huit articles décrivant l'implantation d'un modèle neuronal afin de prédire la performance de filtration ont été recensés (Gagnon et al., 1997; Zhang et Stanley, 1999; Valentin et al., 1999; Yu et al., 2000; Baxter et al., 2001; Baxter et al., 2002; Maier et al., 2004; Hernandez et Le Lann, 2006). Tous ont recours à l'ajout de coagulant chimique, et dans certains cas de polymère ou charbon actif en poudre. Bien que les filtres à sable de la ville de Montréal fonctionnent sans coagulant, ces articles seront étudiés ci-après car ils fournissent des éléments de méthodologie pour l'élaboration un modèle prédictif de la performance de filtration.

La plupart de ces articles décrivent des chaînes de traitement ayant pour procédés de traitement les unités suivantes : coagulation, (floculation) décantation, et filtration. L'ajout d'un coagulant chimique permet d'augmenter la performance de filtration en créant de gros flocs plus facilement décantables. Un dosage optimal de coagulant est nécessaire car une dose trop faible donnera une faible qualité de l'eau en sortie du procédé (et peut être le non respect des normes de production d'eau potable), alors qu'une trop forte dose coûte cher en produit chimique et une grande quantité

d'aluminium résiduel (si l'alun est le coagulant employé) présente un risque pour la santé publique. Usuellement, l'ajustement de la dose de coagulant est effectué manuellement et périodiquement en fonction de l'expérience des opérateurs et de tests en laboratoires (jar-tests). Ces méthodes excluent tout procédé de contrôle en temps réel, c'est pourquoi un modèle de type RNA permettrait d'automatiser le pilotage de l'unité de filtration afin de réduire les coûts d'opération de la station.

Deux types de modèles

Dans le but d'automatiser le contrôle de la performance de la filtration, ces articles émettent la distinction entre deux types de modèles : les modèles de procédés, et les modèles inverses des procédés.

Les modèles de procédés visent à décrire la loi entrée/sortie du phénomène. Un ou plusieurs paramètres de qualité de l'eau à la sortie du procédé sont prédits en fonction de la qualité de l'eau à l'entrée et des paramètres de contrôle de la station (vitesse de filtration, dose de produit chimique, etc.). La plupart des modèles prédisant la turbidité en sortie du décanteur ont pour entrées :

- Les variables de qualité de l'eau brute :
    - Le pH : c'est un facteur essentiel à la coagulation. Les coagulants voient leur efficacité varier selon les plages de pH considérées.
    - L'alcalinité : elle est nécessaire à la réaction de la plupart des coagulants et agit en tant que tampon pour limiter la variation de pH. L'ajout de coagulant consomme de l'alcalinité. Sa valeur est contrôlée dans la suite du traitement afin de limiter la corrosion dans le réseau de distribution.
    - La température : à basse température, les cinétiques chimiques sont ralenties et la viscosité de l'eau augmente.

- La turbidité ou compte de particules : ceci représente la valeur initialement présente de matières solides à enlever. Trop peu ou trop de turbidité sont difficiles à traiter.
- La couleur, l'absorbance UV à 254 nm (UVA-254nm), ou la Demande Chimique en Oxygène (DCO). Ces variables reflètent généralement la quantité de Matière Organique Naturelle (MON) présente. La quantité de MON présente est fortement reliée à la dose d'alun requise.
- La conductivité (plus rarement) : ce paramètre indique l'état de minéralisation de l'eau, c'est peut être un indicateur de qualité de l'eau. En effet, si la source d'eau est le résultat du mélange de deux masses d'eau de qualité et de conductivité différentes, la variation de la conductivité peut être représentative du rapport de mélange entre les sources.
- La teneur en oxygène dissous. Seulement dans un article où le procédé de traitement utilise deux étages de filtres à charbon actif granulaire couplés à une pré et une inter ozonation (Valentin et al., 1999). La teneur en oxygène dissous a une influence sur l'enlèvement supplémentaire qui peut être réalisé par filtration biologique.

- Les variables de contrôle du procédé comme la dose de coagulant, la vitesse de filtration ou le débit de l'usine; et, selon les procédés adoptés, la dose de polymère, et/ou la dose de charbon actif en poudre.

Étant donné que les variables à modéliser sont issues de séries temporelles, certains auteurs utilisent aussi en entrée des variables de qualité de l'eau décalées temporellement, ou bien ajoutent en entrée la variation du paramètre de qualité (variation depuis le pas de temps précédent).

D'autre part, les modèles inverses de procédés ont pour but d'estimer la valeur optimale des paramètres de contrôle en fonction des variables de qualité de l'eau à l'entrée (citées ci-dessus), et des valeurs cibles de la qualité de l'eau à la sortie (par exemple, la turbidité à l'eau décantée ou filtrée).

L'acquisition de données pour les modèles inverses de procédés requiert des essais effectués en laboratoire ou en essai pilote. En usine, il est difficile d'expérimenter la variation des paramètres de réglages dans une large gamme pendant la production d'eau potable ! Une telle base de données a été produite en pilote à Edmonton (AB) (Baxter et al., 2002). En même temps que la qualité de l'eau brute variait au fil des jours, un algorithme faisait varier aléatoirement la dose d'alun injectée afin d'obtenir des données couvrant un large éventail de combinaisons possibles. Il est même recommandé d'utiliser en entrée plusieurs variables cibles à l'eau filtrée afin d'augmenter la plage de fonctionnement du modèle et de gagner en flexibilité d'opération (Maier et al., 2004).

L'ensemble des entrées et sorties retenues par les auteurs cités ci-dessus est récapitulé au Tableau 1-1. Les modèles de procédé et leurs inverses y figurent, les entrées (I) et sorties (O) sont alors interverties.

Tableau 1-1 : Tableau récapitulatif des entrées et sorties retenues des modèles prédictifs de la performance de filtration

	Qualité de l'eau brute											Paramètres d'opération							
Auteur, année	pH	Température	Alcalinité	Turbidité	Compte de particules	Couleur	UVA-254nm	DCO	Conductivité	Dureté totale	Teneur en oxygène dissous	Dose de coagulant	Dose de polymère	Dose de PAC	Débit de l'usine	Débit en sortie du décanteur ou filtre	Turbidité en sortie du décanteur ou filtre	Couleur, eau filtrée	UVA, eau filtrée
Gagnon, 1997	I	I		I					I			O							
Zhang, 1999	I	I	I+ΔI	I+ΔI		I+ΔI						I		I		I	O		
	I	I	I+ΔI	I+ΔI		I+ΔI						O		I		I	I		
Valentin, 1999	I	I		I			I		I		I	O							
Yu, 2000	I			I					I			O					I		
Baxter, 2001	Idem Zhang, 1999																		
Baxter, 2002	I	I	I		I	I						O	O		O	I	I		
Maier, 2004	I		I	I		I	I	I				I					O	O	O
	I		I	I		I	I	I				O						I	I
Hernandez, 2006	I	I		I						I		O							

I : input O : output ΔI : entrée décalée temporellement

Informations de modélisation

Certaines informations utiles en termes de méthodologie sont présentées ci-après.

Tout d'abord, quel pas de temps faut-il adopter ? Les travaux publiés ci-dessus utilisent comme pas de temps : 5 minutes, 15 minutes, ou la moyenne journalière. Le choix du pas de temps est grandement dépendant de la vitesse de variation de la qualité de l'eau brute. Plus elle varie rapidement, plus le pas de temps doit être court. Par exemple, dans la ville de Taipei (Taiwan), lors de la saison des pluies, la turbidité

à l'eau brute passe de 10 UTN à 100 UTN en quelques minutes, l'auteur a donc opté pour un échantillonnage toutes les 5 minutes (Yu et al., 2000).

Par ailleurs, concernant la sélection des exemples pertinents, les données en régime stationnaire sont éliminées afin de réduire le nombre d'exemples redondants et d'empêcher le réseau d'apprendre un comportement récurrent donné. Ainsi, à Sainte-Foy (QC), le modèle de prédiction de la dose optimale de coagulant ne se servit que des mesures autour des périodes transitoires (Gagnon et al., 1997).

Implantation en station

Trois articles abordent l'implantation en station des modèles RNA afin d'assurer une gestion automatisée de la production d'eau ou de fournir un simulateur virtuel pour la formation des opérateurs (Zhang et Stanley, 1999; Baxter et al., 2002; Maier et al., 2004).

Un modèle additionnel prédisant les pH et l'aluminium résiduel à l'eau filtrée en fonction de la qualité de l'eau brute et des doses injectées peut être couplé au modèle de prédiction de la dose optimale de coagulant (Maier et al., 2004).

À Edmonton (AB), les réseaux de neurones furent générés avec le logiciel Statistica®, le code fut extrait en langage Visual Basic®, et une interface graphique fut construite sous Excel®. La version finale de l'interface agit en optimisant le coût d'opération sous contraintes de la qualité de l'eau produite (respect des normes), et contraintes d'opération sur les variables de contrôle. Le logiciel teste toutes les combinaisons des trois variables de contrôle existante (dose d'alun, de polymère, et débit de production d'eau à la station), et préconise la meilleure solution trouvée (Baxter et al., 2002).

Chapitre 2 MATÉRIEL ET MÉTHODES

Le présent chapitre définit la méthodologie retenue pour bâtir les modèles prédictifs. Certaines étapes seront illustrées par les exemples issus du Chapitre 3.

2.1 Récapitulatif des étapes à suivre

Les étapes de développement du modèle comprennent :

1. La définition des objectifs.
2. Le choix du type de modèle (linéaire, RNA, etc.), lequel est le mieux adapté ?
3. La récupération de la base de données par connaissance préalable sur le sujet, et/ou disponibilité des données.
4. Le tri de la base.
5. Le choix des variables d'entrées.
6. Le partitionnement des exemples.
7. Le choix d'une architecture de réseau et des paramètres internes (prétraitement, paramètres internes, algorithme d'apprentissage, fonction d'erreur à minimiser, etc.).
8. La calibration des réseaux (phase d'apprentissage).
9. Le choix d'un critère de performance.
10. La détermination du meilleur réseau.
11. L'assemblage du modèle final.

2.2 Matériel

Les simulations de ce projet ont été réalisées à l'aide du logiciel commercial Statistica version 7.1 et de son module « *Neural Networks* » associé (Statsoft Inc., Tulsa, Oklahoma). De plus, des macro-instructions programmées en Visual Basic ont permis de programmer les boucles nécessaires à la recherche heuristique de solutions. Les ordinateurs utilisés furent équipés de micro processeurs Pentium 4, 3GHz, avec 1Go de mémoire RAM.

2.3 Méthodologie employée

Cette section de méthodologie a été écrite comme un guide de conception à usage de la Ville. Si les modèles développés doivent être implémentés en station, il sera nécessaire de les ré-entraîner périodiquement. Il est ainsi utile d'énumérer étape par étape les hypothèses et les problèmes rencontrés. Une illustration complète accompagnée de valeurs numériques et figures est disponible au Chapitre 3.

2.3.1 Définition des besoins

Le but du modèle est de pouvoir prédire la turbidité à la prise d'eau brute pour une journée à l'avance. Ce modèle doit pouvoir être implanté en station. Les usagers ciblés sont les ingénieurs d'exploitation et les opérateurs, l'utilisation du modèle doit être simple et conviviale.

Ses prédictions doivent être précises et fiables, en amplitude et dans le temps. C'est-à-dire que la valeur numérique produite par le modèle doit approcher la valeur réelle, mais la prédiction temporelle doit être assez bonne pour ne pas manquer le début d'un pic de turbidité. Par exemple, prédire 10 UTN au lieu des 12 observés est plus acceptable que de prédire exactement 12 UTN, mais avec une journée de retard.

La fiabilité s'exprime par la confiance que peut avoir l'opérateur dans le modèle : un modèle générant beaucoup de faux positifs (donc de fausses alertes) pourrait être jugé trop sensible et les prédictions de pointe de turbidité ne seraient pas prises au sérieux.

Nos modèles se doivent d'être robustes. Si, une fois implantés en station, nous décidons après quelques années d'opération de les ré-entraîner avec les nouvelles données disponibles (pour améliorer la performance et/ou tenir compte de la dérive dans le temps de certaines valeurs numériques), le modèle qui fut auparavant une solution optimale, se doit d'être encore performant. On éliminera ainsi les modèles dont les résultats sont extrêmement variables.

2.3.2 Choix du type de modèle le mieux adapté

Réseaux de neurones ou autre ?

Ici, notre objectif est de modéliser la qualité de la prise d'eau brute dans le fleuve Saint-Laurent. Ceci représente un système non linéaire (les mêmes causes ne produisant pas les mêmes effets), complexe et de grande échelle. Les RNA sont particulièrement adapté pour la modélisation non linéaire (Haykin, 1999). De plus, nous pouvons ajouter que notre but est de pouvoir bâtir un modèle rapidement opérationnel pour la ville, c'est pourquoi la modélisation de type « boîte noire » des réseaux de neurones est beaucoup plus appropriée pour nos besoins que l'élaboration d'un modèle conceptuel à l'échelle du Saint-Laurent. Ce modèle conceptuel exigerait la connaissance complète des phénomènes physiques, chimiques, et biologiques ayant lieu dans le fleuve. La taille de la zone d'étude est telle qu'un modèle conceptuel demanderait une puissance de calcul élevée dont la Ville n'a pas utilité (en termes d'investissement).

De plus, les réseaux de neurones artificiels (RNA) sont particulièrement adaptés pour les modèles non linéaires de part leur construction même. L'usage des RNA en hydrologie est bien documenté (voir la revue de littérature au chapitre précédent), et les travaux prometteurs de Tremblay (2004) nous fournissent une bonne base pour démarrer notre projet. Ces raisons expliquent l'utilisation des RNA à la place d'autre types de modélisations (ARMAX, régression linéaire ou polynomiale, etc.)

Type de sortie pour notre modèle : classification ou régression ?

Un modèle de régression permettant de prédire la valeur numérique de la turbidité le lendemain conviendrait en tant qu'outil de gestion pour la quantification des dépassements potentiels aux normes du RQEP. D'un point de vue opérationnel, les deux modèles pourraient convenir car la valeur issue de la régression pourrait par

exemple servir de base au dosage de coagulant, alors que l'établissement d'un dépassement d'un seuil de turbidité par rapport à une valeur jugée « critique » déclencherait une alarme pour les opérateurs qui se prépareraient à une situation d'urgence.

Il est intéressant de noter que les performances de la régression seront moindres qu'un modèle de classification spécifique à un seuil de turbidité donné. Ce dernier n'ayant qu'à se concentrer sur l'équation frontière entre deux classes au lieu de modéliser des valeurs continues.

Nous déciderons donc de coupler notre modèle de régression à des modèles de classification spécifiques à divers seuils de turbidité. Nous pourrons vérifier la concordance des prédictions entre modèles, et ceci résultera en un gain de performance (Tremblay, 2004).

2.3.3 Récupération de la base de données

Comme nous l'avons vu au chapitre précédent, les connaissances préalables obtenues grâce aux travaux de Tremblay (2004) ont permis d'isoler huit facteurs explicatifs des évènements turbides. Les variables à récupérer sont des indicateurs représentatifs de ces phénomènes. Nous avons sélectionné une liste agrandie de ces variables appartenant à trois grandes familles de données : météorologique, hydrologique (débits des principaux affluents et des tributaires secondaires) et qualité de l'eau en amont. Un total de 40 variables constitue la base de données préliminaire (Tableau 3-1).

La zone d'étude est délimitée par le triangle formé par les lacs Saint-François, des Deux Montagnes et la prise d'eau brute à Lasalle.

Nous avons considéré dix années de données au lieu des trois ans et demi utilisés précédemment, ceci afin d'étendre le nombre d'exemples turbides disponibles pour la

calibration de notre modèle. Les données furent recueillies pour la période du 1er janvier 1996 au 31 mai 2006.

La Figure 2-1 présente une carte de la zone d'étude avec la localisation des données disponibles.

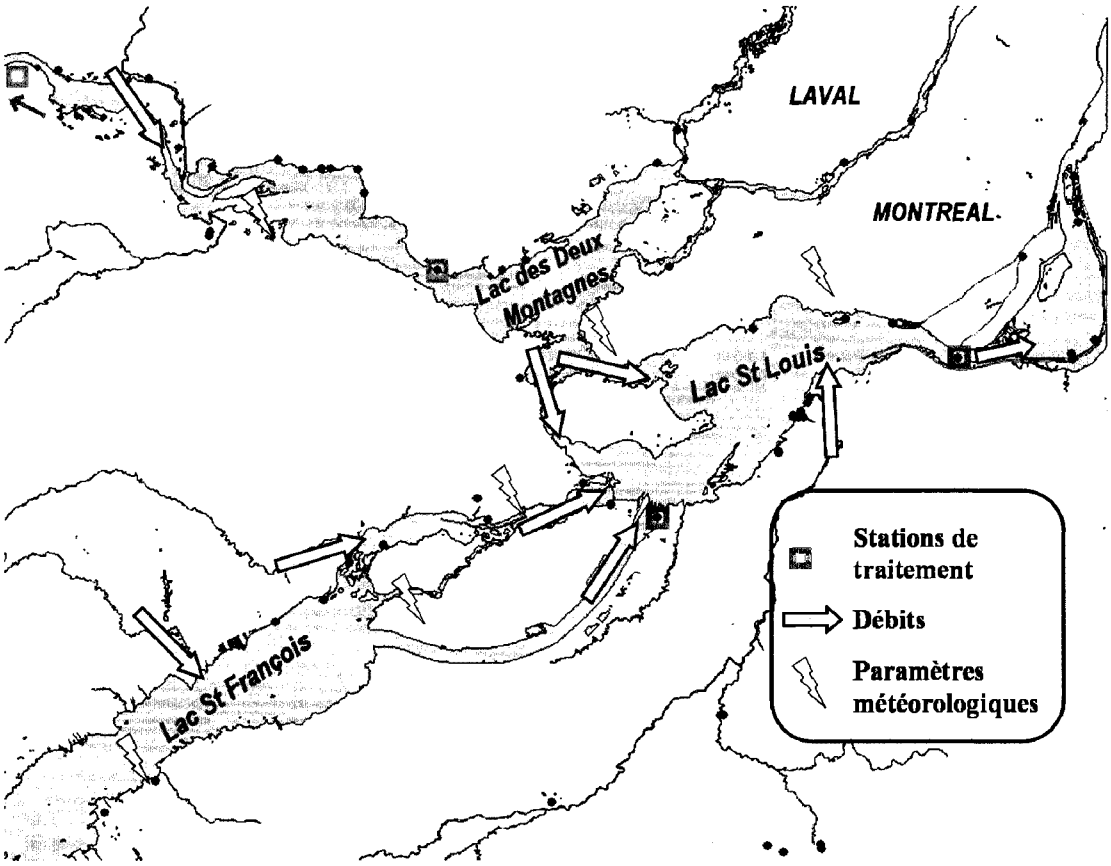

Figure 2-1 : Carte de la zone étudiée, localisation des données

2.3.4 Tri de la base de données

Dans un premier temps, la sortie du modèle est examinée pour observer des particularités comme une périodicité, une saisonnalité, etc. Ensuite, un double tri de la

base de données est effectué : d'abord le tri des variables pertinentes, puis parmi les variables retenues, un tri des exemples pertinents.

Analyse du phénomène à modéliser

Avant d'entamer le tri de la base de données, nous devons fixer une série de définitions afin de caractériser nos phénomènes. Ensuite un examen préalable de la variable de sortie à modéliser peut être fort utile pour orienter nos connaissances vers des comportements spécifiques. Par exemple, dans notre cas, la présence de pointes durant le printemps nous incite à observer les variables représentatives des causes printanières (fonte des neiges, renversement, tempêtes, etc.).

Les définitions à adopter doivent permettre de caractériser les données de turbidité. Une analyse statistique ou la connaissance d'une valeur seuil critique pour l'opération peut définir ce que l'on appelle « un évènement turbide ». Un événement a été jugé « turbide » si sa valeur excède 3.1 UTN (90^e centile). De même, des classes ou catégories de turbidité sont fixées pour discrétiser nos valeurs et en simplifier l'analyse (nous créons cinq groupes : classes I à V). Par ailleurs, nous émettons une distinction entre l'amplitude de l'évènement et sa durée (Tremblay, 2004). Un événement « turbide » est dit « de fond » si sa durée excède cinq jours. Il est dit « superposé » si un pic apparaît durant un événement de fond, il est considéré « isolé » sinon (voir Section 3.1.1).

Ensuite, en traçant la turbidité à l'eau brute en fonction du temps (en journée julienne) pour les dix années de données superposées, une vue d'ensemble du phénomène à modéliser apparaît. La présence de caractère saisonnier nous pousse à explorer des sous-modèles (un par saison), pour mieux identifier les causes explicatives propres à chaque saison, d'où un gain de performance attendu. Dans le cas d'évènements saisonniers, nous pouvons examiner si les pointes de turbidités observées sont les mêmes d'une saison à l'autre (vis-à-vis des définitions adoptées), si ce n'est pas le cas

il y a de fortes chances que différentes causes soient responsables de différents évènements. Une analyse statistique pour l'année et saisonnière vient corroborer ces résultats et justifierait le découpage en saisons. Sur la Figure 3-1, deux saisons principales se distinguent : le printemps et l'automne. Des événements superposés dominent au printemps, alors qu'à l'automne les évènements isolés sont plus nombreux.

Notons que si nous disposons d'un nombre insuffisant de données, ou que les causes explicatives (donc les variables d'entrées) pour chaque saison ne varient pas, alors un modèle unique pour toute l'année pourrait être développé.

Bien que cela ne soit pas le cas dans la présente étude, la présence de périodes, comme les marées, est aussi à envisager car elle peut conditionner le pas de temps, la récurrence du phénomène et le réseau peut être construit pour s'orienter vers la reconnaissance d'une solution périodique (Beaudeau et al., 2001). Si des cycles annuels ou saisonniers sont observés, des entrées reflétant ces cycles peuvent être proposés au modèle (Nour et al., 2006b). Ainsi, les études antérieures comprenaient une variable d'index de saison pour distinguer, l'automne, le printemps et le reste de l'année (Tremblay, 2004).

Cette analyse préliminaire menée sur la sortie du modèle permet d'identifier de façon préliminaire les variables d'entrée. Il convient ensuite d'aller trier parmi toutes les données récupérées celles qui sont pertinentes et d'écarter les autres.

Sélection des variables candidates

Parmi les données disponibles, certaines variables peuvent véhiculer une information utile pour la modélisation (par exemple, la qualité de l'eau directement en amont), mais elles peuvent aussi comporter des limitations conduisant à les rejeter lors de l'élaboration du modèle. Nous allons voir quels sont les critères que doivent remplir

nos variables pour pouvoir être considérées comme acceptables ou rejetées. Trois critères sont retenus : la disponibilité, l'accessibilité et la fiabilité des données.

Il y a tout d'abord la disponibilité des données. Existe-t-il un capteur à l'endroit que l'on souhaite ? Si oui, était-il en fonction durant toute la période où nous recherchons notre base de données ? L'exemple le plus flagrant est la présence de données de qualité de l'eau à Beauharnois de 2001 à 2006 : si nous considérons cette variable incomplète bien que pertinente en termes d'informations, il nous faudra négliger les exemples apportés de 1996 à 2001 (soit la moitié de nos données), notamment les pointes de turbidité de 1998 qui furent assez exceptionnels.

Ensuite, il faut tenir compte de l'accessibilité des données. Y a-t-il coopération de la part des stations ou organismes disposant de telles données ?

De plus, ces données doivent être fiables. La valeur indiquée par le capteur est-elle représentative de la réalité ? Mais surtout, cette valeur a-t-elle le même comportement durant toute la période des mesures ? Par exemple, le déplacement d'un capteur ou le changement de son environnement peut faire translater ces mesures. Les valeurs de turbidité à l'eau brute peuvent être influencées par le point de prélèvement de l'eau acheminée au capteur (phénomènes de décantation). Ensuite, les valeurs rapportées par le capteur peuvent être biaisées, certains capteurs peuvent saturer : par exemple, un capteur de couleur limité à 50 UCV. Au printemps, lors de la fonte et de la dégradation progressive de la qualité de l'eau de la rivière des Outaouais, le capteur de couleur indiquait une hausse jusqu'à la valeur 50 UCV, puis sur dix années printanières observées, au lieu d'indiquer des valeurs supérieures à 50 UCV, voici les valeurs qui ont été recensées : « 0 », « >50 », et « - ». Il semble évident que les valeurs 0 sont erronées, qu'il s'agit réellement d'un nombre supérieur à 50 UCV, et que la valeur « - » ne peut donc être utilisée telle quelle dans notre modèle. L'observation pertinente du phénomène permet de dire que ce capteur n'est pas fiable pour utiliser ces données brutes, il conviendrait de les traiter une par une. Si le nombre d'exemple à

traiter est trop grand vis-à-vis de l'information potentielle supplémentaire que la variable peut apporter, l'abandon de la variable en question est préférable.

Ces trois critères doivent être remplis au moment de la récupération de la base de données, mais aussi dans le futur. Si l'objectif final est d'implémenter le modèle en station, il serait aussi souhaitable de pouvoir récupérer facilement au jour le jour (internet, télémétrie, etc.) des données fiables durant toute la durée de vie du modèle.

Finalement, il faut aussi tenir compte de la redondance de l'information. Des variables comme la couleur et la turbidité à l'eau brute sont fortement corrélées : une seule des deux variables suffirait à apporter l'information nécessaire.

Une fois les variables acceptées ou rejetées (analyse verticale), il convient de regarder pour chacune d'elles les exemples afin de détecter des aberrations (analyse horizontale).

Élimination des exemples non pertinents

Une analyse graphique et une analyse statistique pour chaque variable permet d'identifier des valeurs ou périodes à rejeter. Ceci permet de déceler rapidement des valeurs aberrantes ou manquantes : il s'agit par exemple de discontinuité dans des courbes normalement continues (la température de l'eau brute stagnant à 1°C pendant l'hiver pour subitement passer à 8 °C), ou bien de valeurs aberrantes reflétant un disfonctionnement ponctuel du capteur (la valeur -9999 pour remplacer une valeur manquante dans les systèmes d'acquisition automatique). Il faudrait, dans certains cas, écarter l'exemple en question. L'analyse statistique produit des distributions des exemples et détecte ceux qui sont extrêmes ou aberrants. Certaines de ces valeurs peuvent refléter un disfonctionnement du capteur et/ou un phénomène extrême (une très forte tempête sans précédent historique par exemple). Comme nous nous intéressons à la modélisation de phénomènes extrêmes, nous n'avons pas écarté ces

exemples particuliers. Une expertise couplée à la connaissance historique des situations exactes de l'époque permettrait de trier les exemples pertinents de ceux traduisant un disfonctionnement des capteurs ou des mesures. Or, une telle expertise sur une période de temps très longue ne peut être apportée pour toutes les données récupérées auprès de divers organismes. Cependant, cette connaissance des exemples formant la base de données peut être utile par la suite lors de l'analyse au cas par cas des mauvaises classifications des modèles.

Ainsi, l'étape de tri de la base de données a permis : de convertir les données provenant de différentes sources dans un format exploitable, d'éliminer les variables incomplètes, et parmi les variables restantes, d'éliminer les exemples qui ne représenteraient pas le phénomène à modéliser.

2.3.5 Sélection des entrées du modèle

Une fois le tri préliminaire effectué, le choix des variables est effectué par des méthodes graphiques et statistiques.

Connaissances préalables des causes et variables explicatives

Afin d'effectuer une analyse grossière dans un premier temps, et au même titre que nous avions discrétisé la variable de sortie en classes, les variables seront regroupées en grandes catégories (pluie, vent, hausse des tributaires secondaires, qualité de l'eau en amont, etc.); et, dans certains cas, des variables d'index reflétant une cause explicative donnée seront créées. Par exemple, il s'agit d'index indicateurs de renversement ou de fonte des neiges / fragilité du couvert de glace dont les détails de construction figurent à l'Annexe A.

Les causes explicatives sont ensuite résumées dans un tableau où figurent la cause, les variables explicatives associées, et la valeur seuil considérée pour l'activer. Ce tableau

sera utile par la suite lors de l'analyse graphique des évènements turbides (un exemple est fourni au Tableau 3-5 de l'Annexe B).

Plusieurs phénomènes décrits sont liés. À titre d'exemple, la hausse des tributaires secondaires est une conséquence de fortes pluies (l'été ou à l'automne), ou de la fonte des neiges au printemps. De même, la fonte est associée à une augmentation du débit de passage au barrage de Carillon et donc à un accroissement de la contribution des Outaouais au mélange dans la gire. Il y a donc redondance de certaines variables, certaines pouvant même expliquer plusieurs 'causes'.

Recensement des évènements turbides et analyse graphique

Une fois les valeurs seuils fixées pour caractériser une situation qualifiée de « turbide », un recensement de tous les évènements turbides disponibles pour chaque classe de turbidité peut être effectué : 213 évènements turbides ont été trouvés pour l'usine Des Baillets.

À chaque pic de turbidité sont recensés : la date, l'amplitude du pic, sa durée, la classe à laquelle le pic appartient, mais surtout quelles ont été les variables et causes explicatives « activées » dans les jours qui ont précédés, et combien de jours séparent la cause potentielle de l'évènement turbide. Un exemple de ce type de tableau résumé des causes activées est donné au Tableau 2-1.

Ce recensement fournit une double information : le pourcentage d'occurrence de telle ou telle cause selon la saison, ainsi qu'une appréciation sur la fenêtre temporelle à considérer par variable (voir la section ci-dessous).

Tableau 2-1 : Exemple de tableau récapitulatif du recensement des évènements turbides

Évènement turbide numéro	Date de la pointe	TURB_DB max (UTN)	Classe max	Durée (jours)	Type d'évènement	Vent	Pluie	Renversement	Fragilisation du couvert	Fonte des neiges	Hausse des tributaires	Contribution Outaouais
95-a1	24-01-1996	3,6	3	2	Isolé		X		X		X	
96-p1	20-04-1996	3,6	3	1	Isolé		X	X				
96-p2	23-04-1996	16	5	12	Fond	X	X	X	X	X		X
96-p3	24-04-1996	5,9	4	1	osé	X	X		X	X		X

Décalages temporaux à considérer

La sélection de variables retenues initialement peut aussi inclure les mêmes variables avec les décalages temporaux appropriées. En effet, la modélisation d'un signal issu d'une série temporelle par un réseau non récurrent (modèle dit statique ou non dynamique) nécessite d'inclure des entrées supplémentaires pour pallier à cette difficulté. Ces entrées additionnelles sont une sélection des variables d'entrées pertinentes retardées de x jours. Le nombre de jours au-delà duquel la variable décalée n'a plus d'influence sur la sortie est appelé « fenêtre temporelle ». Elle peut être déterminée par : une connaissance préalable sur le sujet (temps de séjour de l'eau entre deux points par exemple), les résultats de modélisations antérieures, ou devinée grâce à des analyses graphiques ou statistiques. La première (i.e. analyse visuelle) est faite lors du recensement ci-dessus, la deuxième est réalisée avec les méthodes décrites ci-après.

Analyse statistique – corrélation

Cette analyse est effectuée avec le tracé des diagrammes d'auto-corrélation et de corrélation croisée entre la variable et la sortie en fonction du décalage temporel (dont les schémas types sont représentés à la Figure 2-2).

Tout d'abord, il est important de rappeler que ces méthodes ne peuvent fournir que la composante linéaire entre la variable et elle-même ou avec la sortie. Étant donné que le phénomène à modéliser est non linéaire, il se peut qu'une variable peu corrélée avec la sortie puisse être utilisée pour la prédiction (à titre d'illustration les variables retenues pour le modèle final avaient des coefficients de corrélation croisée de l'ordre de 0,8 à 0,3). Il faudra donc oublier la valeur usuelle de 0,85, symbolisant généralement une bonne corrélation en modélisation, et ne considérer les résultats donnés par ces courbes que comme des indications relatives.

Par ailleurs, l'auto-corrélation est généralement une fonction strictement décroissante : les variables des jours précédents ayant de moins en moins de lien avec la variable à la date t_0. Une brusque chute suivie d'un replat nous indique qu'au-delà du temps t_1 les décalages ne sont pas liés linéairement à $I_n(t_0)$: pour tout $t \geq t_1$, $I_n(t_0\text{-}t)$ n'est plus corrélé avec $I_n(t_0)$. Ce graphe donne une indication du décalage temporel à considérer sur une variable pour limiter la redondance des informations (par exemple : le débit de la rivière des Raisins reste auto-corrélé jusqu'à 3 jours de la valeur 0.8, alors que la pluie du jour j au suivant n'est pas corrélée).

En ce qui concerne le graphe de corrélation croisée de la variable d'entrée $I_n(t_0)$ avec la sortie $O(t_0)$ en fonction des décalages temporaux de I_n, l'allure générale de la courbe serait une courbe croissante (optionnel si les temps de réponse sont rapides, le graphe est une courbe décroissante), puis décroissante. Dans un premier temps, de 0 à t_2, $I_n(t)$ est faiblement corrélé avec O. Il peut s'agir d'un temps de réaction du phénomène comme par exemple la hausse graduelle de la turbidité un à deux jours après une forte pluie. Cette corrélation croît jusqu'à un maximum à la date $t\text{-}t_3$. Puis l'influence s'estompe jusqu'à tomber dans un « fond de vallée » à partir de l'instant t_4. Ce décalage t_4 pourrait représenter la fenêtre temporelle. Ce graphe nous permet d'estimer la valeur de la fenêtre temporelle ainsi que le décalage où I_n sera le plus corrélé linéairement avec la sortie $O_n(t_0)$ (soit la date $t\text{-}t_3$).

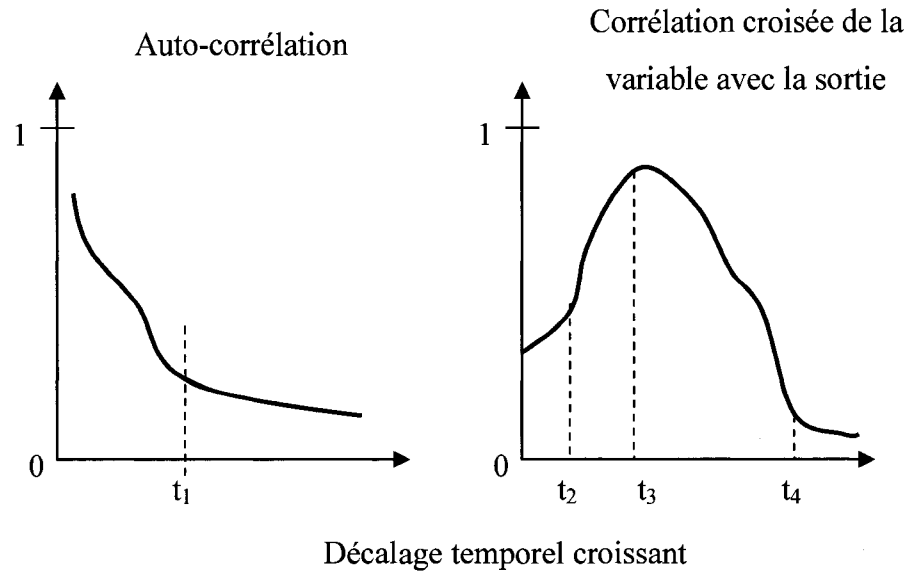

Figure 2-2 : Exemples d'auto-corrélation et de corrélation croisée de la variable d'entrée candidate I_n en fonction du décalage temporel

Exemple : les 40 variables disponibles initialement pour la prédiction de la turbidité à Des Baillets sont réduites à 26 variables potentielles une fois les entrées incomplètes éliminées, et une fois les entrées redondantes regroupées. Après détermination des fenêtres temporelles, ces 26 variables passent à 91 variables potentielles en entrées des modèles. Il est nécessaire de réduire le nombre d'entrées potentielles à l'aide des méthodes suivantes.

Analyse statistique – séparation linéaire évidente ?

L'analyse présentée ici n'a pas de valeur en termes de critère de sélection rigoureux, il s'agit juste d'une vérification graphique utile pour la compréhension du système, elle permet d'obtenir une meilleure idée du pouvoir séparateur des variables dans un plan donné. Elle se présente selon deux graphiques.

Par la suite, nous parlerons de chevauchement des classes selon les variables (X_1, X_2) lorsqu'une ou plusieurs données appartenant à une classe (C_1) sont entourées de données appartenant à une autre classe (C_2) dans un plan (X_1, X_2). A contrario, les ensembles de données seront dits disjoints.

La première analyse est de type boîtes à moustaches. Nous traçons les diagrammes boîtes à moustaches de chaque variable, diagrammes groupés par classe de turbidité : soit les classes I à III, IV, et V (i.e. respectivement les données inférieures au 95ème centile, du 95ème au 99ème centile, et supérieur au 99ème centile). Ceci nous permet d'avoir un aperçu du chevauchement des valeurs des variables entre les classes : peut-on trouver une valeur de cette variable qui va séparer le 75ème centile de la classe considérée, du 25ème centile de la classe immédiatement au dessus ?

La deuxième analyse est un nuage de points catégorisé par les classes dans un plan formé par deux descripteurs (deux variables d'entrées potentielles). Selon le couple de variables choisies, nous allons obtenir une plus ou moins bonne séparation des classes. Ceci nous donne une idée du pouvoir séparateur de telle ou telle variable projetée dans un plan donné. Ce graphe peut aussi mettre en évidence la présence d'exemples problématiques, i.e. dont la bonne classification obligerait à inclure beaucoup de faux positifs à cause du chevauchement des classes.

À titre illustratif, la Figure 2-3 montre peu de chevauchement observé pour la variable OUT_FLV-1 entre les classes IV et V et celles inférieures, comme en témoignent les 25^{e} et 75^{e} centiles. Pour la classe V, la médiane (P_{50}) dépasse le 75^{e} centile de IV. On peut en conclure qu'avec la variable OUT_FLV-1 (valeur la veille du pourcentage de la rivière des Outaouais par rapport au fleuve Saint-Laurent) la classe IV est bien séparée des classes I, II, ou III. Alors que les classes IV et V se chevauchent.

La Figure 2-4 nous montre qu'il y a malgré tout beaucoup de chevauchement entre les classes et que certains exemples ne pourront être bien classés uniquement par ces deux

descripteurs. En effet, trois exemples de la classe V se retrouvent dissociés de leurs homologues, ils sont dans la partie centrale de la figure mêlés aux exemples de classe I à III. Ces exemples ne pourraient être expliqués seulement par les variables OUT_FLV-1 et RIV_CHAT-3 (débit de la rivière Châteauguay trois jours avant).

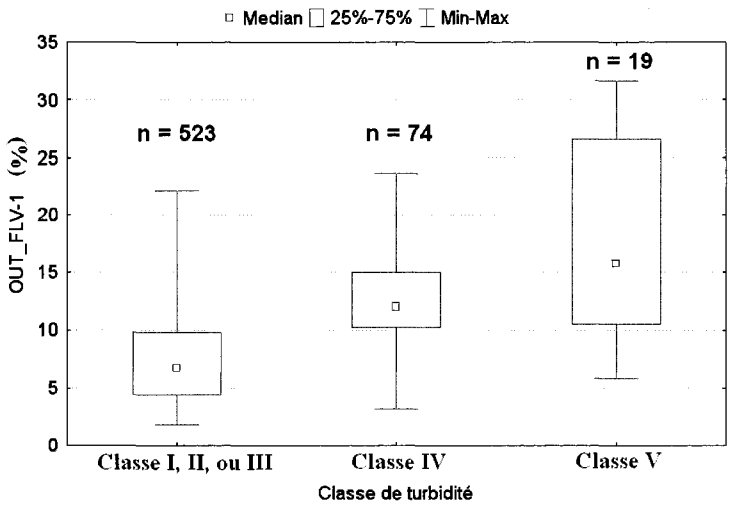

Figure 2-3 : Diagramme boîte à moustaches du pourcentage de la rivière des Outaouais sur le fleuve Saint-Laurent (OUT_FLV-1) par classes de turbidité – printemps

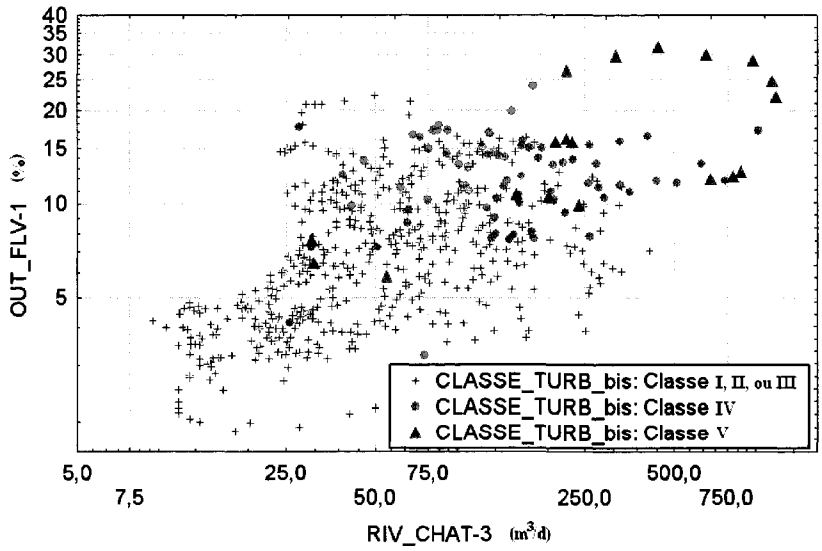

Figure 2-4 : Diagramme nuage de points catégorisé dans un plan (RIV_CHAT-3; OUT_FLV-1) – printemps. RIV_CHAT : débit de la rivière Châteauguay. OUT_FLV : contribution des Outaouais

Analyse statistique – analyse discriminante

L'analyse discriminante a pour objectif de trouver le meilleur sous-ensemble de variables séparant linéairement les classes de turbidité. Parmi un très grand nombre d'entrées potentielles (jusqu'à plusieurs dizaines en même temps), l'algorithme choisit étape par étape la meilleure entrée à ajouter au modèle pour en améliorer la performance de classification. L'analyse s'arrête lorsque toutes les variables ont été incorporées, ou que l'ajout d'entrées supplémentaires n'a plus d'impact significatif sur la performance. Un classement des entrées retenues par ordre d'importance dans la classification est disponible en sortie, d'où un gain de connaissance du système pour le modélisateur.

Bien qu'elle soit linéaire, cette méthode de sélection des entrées nous apporte une solution unique (justement grâce au côté linéaire). L'analyse est robuste car nous pouvons traiter plusieurs dizaines de variables candidates à la fois.

L'évaluation de la performance réelle de chaque classificateur se fait par validation croisée. Les exemples sont séparés aléatoirement en deux ensembles (Train et Test) selon la proportion 80% – 20 % pour chaque classe (basse – haute).

Pour chaque saison, plusieurs analyses discriminantes sont menées (une par seuil classifiant de turbidité), chacune d'elles livrant une sélection de variables d'entrées qui peut différer d'un seuil à l'autre. En observant l'influence de la valeur seuil et des entrées retenues sur le pourcentage de classification correcte, ceci vient confirmer que la création d'une série de classificateurs spécialisés à des seuils de turbidité fut une bonne idée.

Par ailleurs, l'inclusion de deux variables supplémentaires comme la date en journée julienne et la saison parmi les variables potentielles, permet de vérifier que le découpage en saison est adéquat (si ces variables sont bien classées dans la liste des descripteurs influents).

À une saison donnée, les entrées retenues pour le modèle de régression appartiendront à un ensemble issu de l'union des sous-ensembles de variables retenues pour chaque seuil de classification. Ensuite, l'analyse heuristique automatisée Intelligent Problem Solver (IPS) de Statistica déterminera un sous-ensemble optimal de variables pour la régression.

Pour la turbidité à Des Baillets, après analyse discriminante, les entrées potentielles sont réduites de 91 à moins de 25 variables selon les saisons. Certaines des variables retenues sont redondantes, il faut choisir parmi elles lesquelles véhiculent le plus d'information pour l'élaboration du modèle RNA.

Analyse statistique – réseau de neurones probabiliste

Une fois l'analyse de tamisage menée par analyse discriminante, il reste une sélection réduite d'entrées potentielles par seuil et par saison. Pour chaque saison, les ensembles

d'entrées retenues par seuil sont regroupés dans un sur-ensemble résultant de l'union de ces ensembles. Cet ensemble d'entrées potentielles sert de base pour confirmer la sélection d'entrées préconisée pour chaque seuil de turbidité et chaque saison.

Tel que suggéré lors de la revue de littérature, l'utilisation de réseaux de neurones probabilistes (« *probabilistic neural network* » ou PNN en anglais) permet d'inclure les effets non linéaires liant entrées potentielles et sortie du modèle, lors de l'évaluation de la performance de classification d'un groupe d'entrées donné. Ces réseaux présentent l'avantage d'offrir l'unicité de la solution une fois l'apprentissage effectué, contrairement aux réseaux de type perceptron multicouches tributaires de l'initialisation de leurs paramètres libres. La calibration de tels réseaux nécessite le découpage des exemples en trois ensembles (train, select et test) tel que décrit dans la section 2.3.6. Ces réseaux sont couplés à des méthodes d'exploration des combinaisons d'entrées potentielles.

Dans le meilleur des cas, si peu d'entrées sont présentes, la recherche exhaustive parmi toutes les combinaisons possibles peut être menée dans un temps acceptable. Cependant, en général, il faut se contenter d'autres méthodes de recherche étape par étape.

Le recours au module « *feature selection* » de Statistica utilise des algorithmes pré programmés afin de tester des combinaisons d'entrées potentielles et de comparer leur performance de classification à l'aide de réseau de type PNN.

La première est appelée « Stepwise Forward Selection ». À l'étape 0, aucune entrée n'est considérée, l'algorithme entraîne n réseaux à une entrée, n étant le nombre d'entrées potentielles. Le meilleur réseau à une entrée est retenu à la fin de l'étape 1, l'étape 2 inclus d'office cette entrée, et teste sa performance associée à une des n-1 entrées restantes. L'algorithme fonctionne ainsi de suite jusqu'à ce que toutes les entrées aient été incluses ou que l'adjonction d'une entrée supplémentaire ne modifie

plus la performance du modèle. Au contraire, la méthode « *Stepwise Backward Selection* », considère dès le début un réseau doté de toutes les entrées potentielles. À chaque étape, elle élimine l'entrée créant le moins de perturbation sur la performance de classification jusqu'à ce que toutes les variables aient été éliminées. Au final des deux méthodes, le logiciel retient l'ensemble d'entrées ayant donné la meilleure performance. Bien que lourd en termes de calcul dans les premières étapes où toutes les variables d'entrées sont présentes, cette méthode par élimination plutôt que par construction (« *stepwise forward* ») est préférable si peu d'entrées sont à tester (moins de trente). Cette méthode présente de surcroît l'avantage d'être moins sensible aux entrées redondantes (Statsoft, 2006). Elle sera donc privilégiée pour orienter notre choix d'entrées. Il est intéressant de noter l'ordre dans lequel apparaissent ou disparaissent les entrées et la variation relative du critère de performance à chaque modification, ceci aide le modélisateur à comprendre l'importance de chaque entrée.

La deuxième méthode d'exploration utilise les algorithmes génétiques. Dans cette méthode d'optimisation globale, à chaque époque, une population symbolisant des combinaisons aléatoires des entrées potentielles est créée. Les combinaisons les plus performantes suite à l'apprentissage avec le réseau PNN sont 'croisées' entre elles pour donner naissance à une génération fille mêlant les caractéristiques des deux parents. Pour ne pas rester coincé dans un minimum local, une partie de la population est systématiquement générée au hasard pour apporter de la diversité. Au fil des époques, les meilleures combinaisons d'entrées se voient favorisées. À la fin du nombre d'époques prédéterminé, le logiciel retient la meilleure combinaison d'entrées (Statsoft, 2006).

Ces deux méthodes (« *Stepwise Backward* » et algorithmes génétiques) sont menées pour deux échantillonnages distincts afin de s'assurer que les résultats obtenus sont insensibles à la manière dont sont répartis les exemples. Une entrée sera retenue si elle a été proposée par les deux méthodes.

Choix de variables d'entrée potentielles du modèle

Toutes les méthodes citées ci-dessus ont permis d'isoler pour chaque saison et chaque seuil une liste réduite de variables d'entrées du modèle avec les décalages temporaux adéquats.

Pour chaque modèle, d'autres ensembles d'entrées supplémentaires sont testés, à savoir :

- Un sous-ensemble des entrées retenues où sont éliminées les variables jugées redondantes par le modélisateur (par exemple, la pluie le même jour en une autre station météorologique).
- Le sur-ensemble formé de l'union de toutes les variables d'entrées retenues pour la saison considérée.
- L'influence d'un prétraitement alternatif reflétant la distribution propre à chaque variable (voir Annexe D).

En conclusion, l'étape de sélection des entrées a utilisé plusieurs techniques (visuelles et statistiques) pour minimiser la liste des entrées potentielles pour les modèles RNA. À partir de cette sélection réduite, une combinaison optimale d'entrées de chaque modèle pourra être trouvée par essais et erreurs.

2.3.6 Partitionnement des exemples

Les exemples disponibles sont triés en trois grands ensembles : « Train », « Select » et « Test ». Ces sous-ensembles doivent être similaires les uns aux autres : ils doivent contenir sensiblement la même proportion d'évènements de basse et haute turbidité (selon le seuil considéré) et être statistiquement semblables. Les RNA extrapolent mal au-delà des données qui ont servis à leur calibration (Haykin, 2004) : les évènements extrêmes de turbidité doivent donc être laissés à l'ensemble d'apprentissage.

Pour vérifier que les trois ensembles soient représentatifs de la même population, le test non paramétrique de Statistica appelé « analyse de variance de Kruskal Wallis et test de la médiane » est utilisé (Statsoft, 2006). Ce test permet de comparer une série de variables indépendantes classées en trois sous-ensembles et plus. Comme il est non paramétrique les variables n'ont pas à être distribuées normalement. Cinq grandes causes sont considérées au travers des cinq variables suivantes : la turbidité à Des Baillets (TURB_DB), la vitesse moyenne du vent au lac Saint-François (LSF_VITM), les précipitations à Dorval (DOR_PREC), le débit de la rivière Châteauguay (RIV_CHAT), et le pourcentage de contribution de la rivière des Outaouais (OUT_FLV, uniquement au printemps). Ces variables ont été choisies suite à l'analyse discriminante lors de l'étape de sélection des entrées du modèle. Les dernières s'avèrent véhiculer beaucoup d'information pour la prédiction de la première, elles reflètent respectivement le vent, la pluie et la fonte des neiges ou une forte pluie (et la contribution des Outaouais lors de la fonte). Si les écarts donnés par les tests sont faibles avec un « *p value* » élevé nous pouvons considérer que les ensembles ne sont pas statistiquement différents.

À posteriori, il faut aussi vérifier qu'il y ait une proportion d'exemples « bas » sur « haut » sensiblement identique entre les ensembles; de plus, l'observation visuelle des diagrammes boîtes à moustaches pour chaque variable importante à la saison considérée en fonction des trois ensembles de répartition donne une idée de la similitude entre chaque ensemble. Un exemple de ce type de diagramme est fourni à la Figure 2-5.

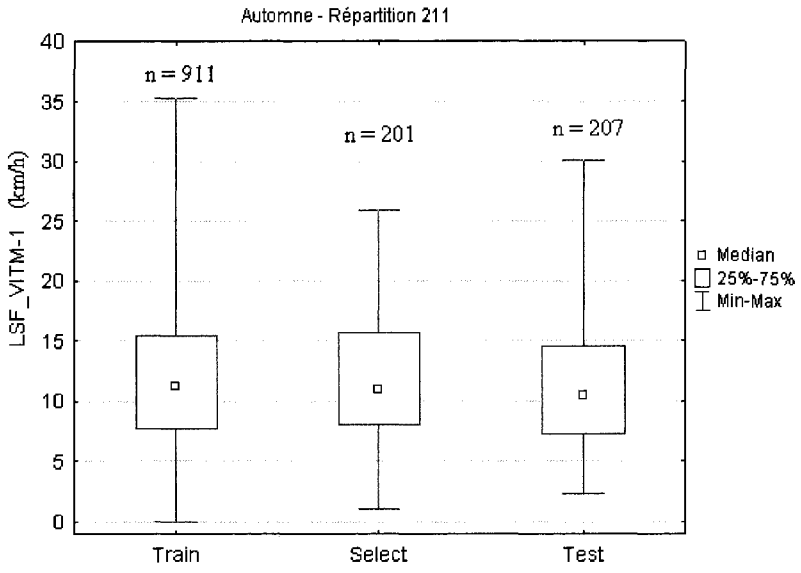

Figure 2-5 : Exemple de vérification boîtes à moustaches du partitionnement pour la vitesse du vent au lac Saint-François la veille (LSF_VITM-1) à l'automne

Afin de vérifier que la performance d'un réseau ne soit pas dépendante du partitionnement utilisé, il est nécessaire de créer pour chaque saison et pour chaque seuil plusieurs échantillonnages différents. Le détail des méthodes d'échantillonnage est donné dans l'Annexe C.

2.3.7 Choix d'une architecture de réseau et des paramètres internes

Architecture considérée : perceptron multicouches

Dans la présente étude, seuls les réseaux de type perceptron multicouche (PMC) à une seule couche cachées de neurones sont considérés. Cette architecture est connue depuis longtemps, bien documentée, robuste, et surtout elle suffit pour approximer numériquement n'importe quelle fonction continue sous réserve d'avoir une complexité suffisante (Haykin, 1999).

Notons, en ce qui concerne les aptitudes de généralisation du modèle, que le PMC à une couche cachée possède des capacités de généralisation supérieures aux réseaux à

base radiale ou aux réseaux de neurones de régression généralisée (respectivement appelés RBF et GRNN en anglais). Ces derniers modélisent la densité de probabilité jointe entre la sortie et les entrées à l'aide de gaussiennes centrées sur les données : leurs prévisions tendent vers zéro 'loin' des exemples vus pendant l'apprentissage. Par conséquent, ils nécessitent un grand nombre de données représentatives du phénomène à modéliser, si possible tous les cas de figures que l'on pourrait rencontrer dans la nature. Pour les raisons citées ci-dessus, la modélisation des concentrations en sédiments dans un réservoir a retenu cette approche pour fournir des prédictions physiquement plausibles, c'est-à-dire pas de concentration négative pour les faibles valeurs (Cigizoglu, 2005). Ce problème de prédictions négatives sera contourné en utilisant une fonction d'activation strictement positive en sortie du modèle de régression (i.e., la fonction logistique). L'utilisation des GRNN pour le modèle opérationnel serait envisageable ultérieurement lorsque plus de données auront été acquises.

Ici, il est important de conserver la capacité de généralisation et d'extrapolation du modèle, pour prendre en compte les dérives éventuelles des données au fil des années de fonctionnement entre deux réapprentissages. Ceci justifie le choix des PMC.

Par expérience, la plupart des articles étudiés lors de la revue de littérature utilisent une seule couche cachée. Une complexité supérieure du modèle pouvant être obtenue avec la deuxième couche, comme la modélisation de frontière non convexe et disjointes (Bishop, 1995), la possibilité d'une deuxième couche sera étudiée seulement si le nombre de neurones de la première couche doit être grand pour obtenir une performance acceptable. En pratique, peu de neurones et une seule couche cachée se révèleront suffisants.

Complexité du modèle, approche par essais et erreurs

Les paramètres jouant sur la complexité du modèle sont les groupes d'entrées choisis et le nombre de neurones de la couche cachée. Le premier, les entrées, dépend surtout de la pertinence des variables retenues; le second, le nombre de neurones, influence directement la complexité du modèle.

Une approche quasi exhaustive est adoptée : à un groupe d'entrées donné, la complexité du modèle augmentera dans un intervalle fixé (i.e. le nombre de neurones de la couche cachée). Parmi tous les réseaux entraînés, nous retiendrons le réseau donnant les meilleures performances avec le plus bas nombre de neurones cachées (voir l'approche multicritères de la Section 2.3.10).

Par exemple, pour la turbidité à l'eau brute à Des Baillets, la plage de neurones considérée est : de un à douze neurones avec un pas de un, de quatorze à 50 neurones avec un pas de deux, et de 54 à 78 neurones avec un pas de quatre.

Paramètres internes

Il s'agit là du prétraitement des variables (voir Annexe D), de l'algorithme d'apprentissage et ses paramètres internes, et de la fonction d'erreur à minimiser lors de l'apprentissage. Tous les détails concernant ces paramètres sont fournis dans l'Annexe F.

Configuration de réseau

Nous appelons « *configuration de réseau* » une combinaison des paramètres : saison, combinaison d'entrées, prétraitement ou non, seuil turbide de classification, répartition des exemples selon la variable d'échantillonnage et nombre de neurones dans la couche cachée.

La recherche heuristique parmi les configurations de réseau de notre plan d'expériences fut menée par l'algorithme Intelligent Problem Solver (IPS) de Statistica. Une macro-instruction, implémentée en langage Visual Basic et utilisant l'analyse IPS, a permis à l'aide de boucles récursives d'explorer l'influence du nombre de neurones de la couche cachée sur la performance obtenue à configuration donnée. La description de cette macro-instruction est donnée dans l'Annexe F.

2.3.8 Calibration des réseaux

Il s'agit de la phase dite d'apprentissage. Un apprentissage de type supervisé fut retenu, c'est-à-dire que les couples (entrées; sortie) servirent pour étalonner les variables libres du réseau (poids des connexions et biais).

À partir du partitionnement des exemples en trois ensembles, l'apprentissage va se faire à partir des données de « Train ». L'ensemble Select est utilisé pour optimiser l'arrêt de l'apprentissage par la méthode de validation croisée.

L'apprentissage est un phénomène probabiliste: selon l'initialisation des poids, l'algorithme d'apprentissage va converger vers divers minimum locaux de la surface d'erreur. Afin de se donner une idée de la variabilité de nos résultats, 40 réseaux à configuration fixée seront entraînés et seulement les dix réseaux les plus performants sur l'ensemble Select seront retenus (selon la fonction d'erreur à minimiser choisie ci-dessus).

La calibration a été menée avec l'Analyse IPS, les détails concernant les algorithmes utilisés pour l'apprentissage et les paramètres associés figurent en Annexe F.

2.3.9 Choix d'un critère de performance

Les réseaux calibrés associent à chaque exemple de turbidité observé une valeur prédite, il faut convertir les nombres obtenus en résultats facilement interprétables

pour juger du réseau répondant le mieux aux besoins de la Ville (i.e. dépendamment des objectifs du réseau). Nous retiendrons la solution donnant de meilleurs résultats, selon nos critères de performance, et étant la plus parcimonieuse (en termes de complexité du réseau), pour pouvoir converser la faculté de généralisation.

Dans l'Annexe G, sont rappelés les besoins auxquels doit répondre le modèle (avec l'emphase sur la nécessité de prédire le moment d'apparition de la pointe de turbidité), il en résulte le détail de l'approche multicritères adoptée ici pour évaluer les résultats des configurations de réseaux sur les plans : performance globale, capacité de généralisation, et variabilité des résultats.

2.3.10 Choix du meilleur réseau

Pour chaque saison, pour chaque seuil de turbidité : un ou plusieurs réseaux semblant remplir le mieux possible les objectifs sont sélectionnés. Il est possible d'agir sur trois paramètres : le groupe d'entrées à considérer, le prétraitement ou non des variables, et le nombre de neurones de la couche cachée.

Description de la méthode

À une configuration donnée, il faut analyser quatre critères de performance et leur variabilité. Pour chaque critère, un intervalle de neurones de la couche cachée donnant les meilleures performances avec les variations les plus faibles est retenu. Les résultats étant variables, et afin que le modèle puisse être toujours performant si ré-entraîné quelques années plus tard, des résultats uniques et absolus ne peuvent être donnés : il s'agira quelques fois de plages optimales de nombre de neurones cachés. À performance égale, l'intervalle de neurones cachés le plus bas sera privilégié afin de conserver la capacité de généralisation du modèle. Pour la construction du modèle final, il faudra rechercher, localement par analyse IPS, une solution optimale dans cet intervalle restreint.

Tous les graphiques qui suivent vont avoir pour abscisse le nombre de neurones de la couche cachée, et en ordonnée, le critère de performance considéré. La démarche s'effectue en six étapes (Figure 2-6) :

1. Pour chaque répartition (appelée A et B), le ou les meilleurs groupes d'entrées (en cas de performance équivalente) sont déterminés. Tous les ensembles d'entrées non prétraitées sont comparés entre eux avec le tracé en nuage de point de la moyenne de chaque critère. Les autres combinaisons sont éliminées.
2. Une liste de neurones cachés donnant une bonne performance pour chaque groupe d'entrées retenus est identifiée. Les diagrammes boîtes à moustaches des quatre critères de performance en fonction du nombre de neurones cachés sont tracés. Ainsi, il est possible d'évaluer la performance globale avec la « matrice de perte somme », puis la performance de généralisation avec les trois autres critères (matrice de perte TEST, matrice de performance TEST et pourcentage de classification correct TEST).
3. Est-ce que cette liste est la même en utilisant une autre répartition ? À priori oui car le phénomène physique modélisé est indépendant de comment sont répartis les exemples. Sinon, il faut agrandir la liste.
4. Pour ces ensembles d'entrées retenus, quel est l'influence du prétraitement? L'étape deux est reconduite avec les données pré traitées par fonction de répartition. On identifie une liste de neurones cachés donnant des résultats optimaux. Ce ne sont a priori pas les mêmes qu'en (2) car les données ont été transformées non linéairement.
5. On compare du groupe d'entrées optimal avec ou sans prétraitement en traçant le diagramme en nuage de points de la moyenne des critères de performance.
6. Choix des réseaux optimaux. Des commentaires sont émis sur l'utilisation du prétraitement. Puis, une liste de neurones cachés à investiguer dans l'élaboration du modèle final est proposée.

Un exemple d'application de cette méthode est fournit dans la section 3.1.8.

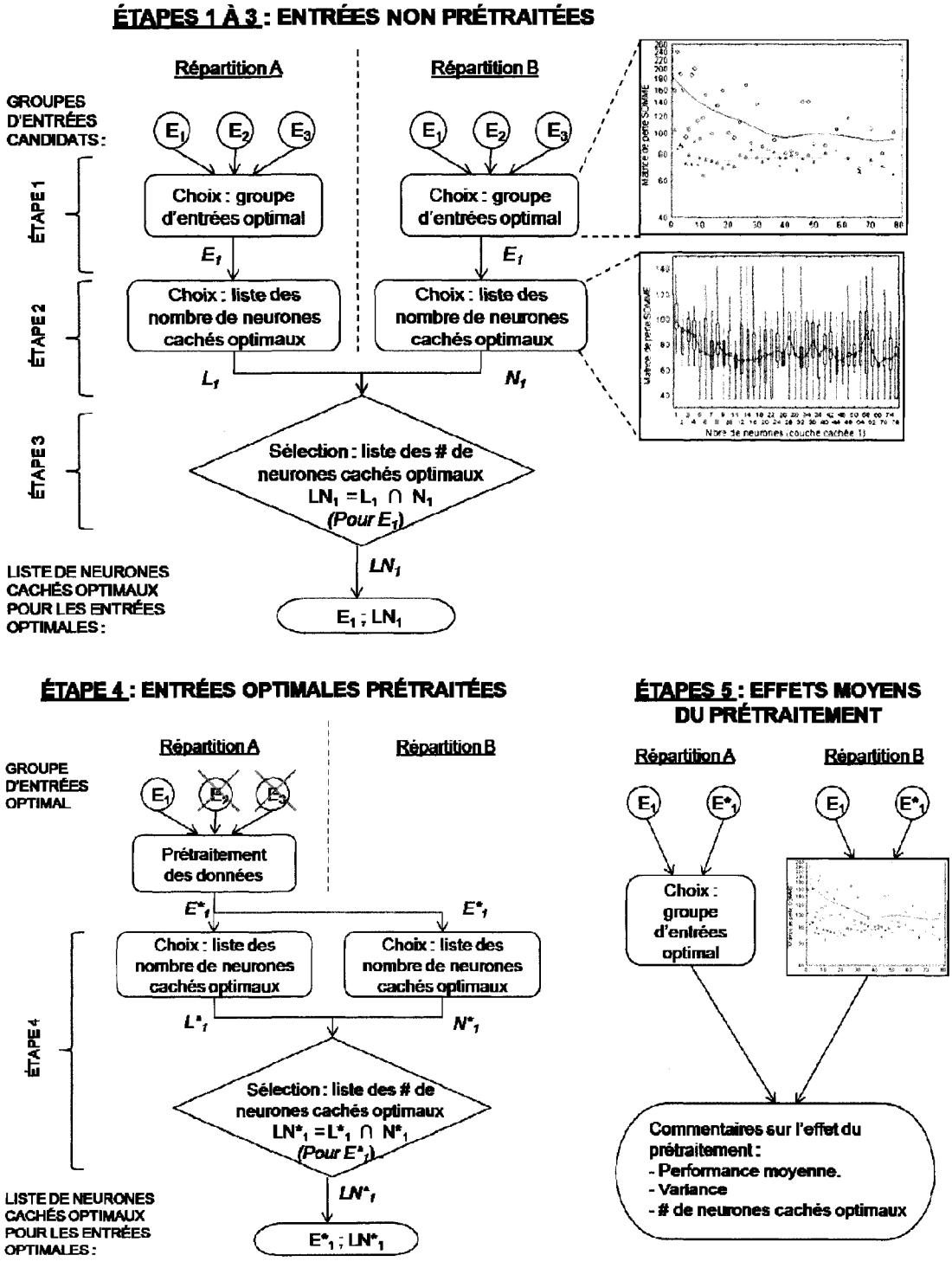

Figure 2-6 : Schéma de la méthode de choix du meilleur réseau

Amélioration des prévisions ?

Les résultats identifiés ci-dessus pourraient être optimisés. En effet, dans l'analyse détaillée des exemples sous classés (faux négatifs), deux situations peuvent se présenter.

Tout d'abord, les faux négatifs peuvent être à la limite du seuil de turbidité considéré, l'équation frontière entre deux classes n'étant pas parfaite, il y aura toujours quelques cas limitrophes mal classés. Par exemple, pour le seuil TURB_DB = 5,5 UTN, des exemples mal classées d'amplitude 5,6 ou 5,8 UTN ressortent. Leur impact sur l'erreur de prédiction commise est faible grâce à l'utilisation de plusieurs modèles de classification bas/haut pour différentes valeurs seuils de turbidité.

Deuxièmement, le mauvais classement est dû à une cause explicative non prise en compte parmi les entrées du modèle neuronal. Il faut alors revenir dans le fichier de recensement pour voir quelles étaient les variables actives pour l'exemple mal classé. Si une variable non prise en compte est identifiée, il convient de l'inclure dans le groupe d'entrées et de refaire les étapes un à six décrites ci-dessus.

Meilleur réseau pour la régression

Une fois les combinaisons d'entrées optimales identifiées pour chaque seuil de turbidité. Pour chaque saison, nous créons un nouvel ensemble d'entrées qui est l'union de ces ensembles. Ce sur-ensemble d'entrées représente toutes les entrées pertinentes pour la classification selon les quatre seuils adoptés, certaines entrées y sont sans doute redondantes, il convient de les éliminer par une des méthodes de sélection des entrées.

Le choix des entrées sera confié au module « feature selection » avec les réseaux de type GRNN de Statistica, ceci afin d'élaguer les entrées non pertinentes (voir le paragraphe sur les réseaux de neurones probabilistes de la section 2.3.5). Puis, le choix

de l'architecture optimale sera confié à l'analyse IPS de Statistica en lui laissant la possibilité de choisir le nombre de neurones cachés, et au besoin un sous-ensemble d'entrées. L'analyse tournera pendant deux jours afin de s'assurer que la recherche heuristique de la solution optimale puisse explorer suffisamment de combinaisons.

Nous retiendrons le réseau offrant la meilleure corrélation avec les données observées, sur l'ensemble de « Test ». Puis, en deuxième critère, ce réseau devra aussi garantir une erreur quadratique moyenne la plus faible possible.

2.3.11 Bâtir le modèle final

Après avoir trouvé une solution optimale pour chaque sous modèle, il est nécessaire de regrouper toutes nos prédictions ensembles afin de fournir un outil prédictif opérationnel tout au long de l'année. En effet, un découpage temporel (en saison) couplé à un découpage de la variable de sortie (en quatre sous modèles de classification plus un modèle de régression) a été effectué pour améliorer les performances de prédiction. Il convient de créer une interface liant les décisions apportées par chaque modèle. Ce problème se décompose en deux étapes.

Mise en commun des modèles – prédiction par saison

Pour chaque saison, les quatre modèles de classification ainsi que le modèle de régression sont regroupés en parallèle pour fournir une prédiction tout au long de la saison. Seront affichés à l'opérateur la synthèse des résultats de classification, plus le résultat du modèle de régression.

Cette synthèse des modèles de classification est calculée avec le résultat des modèles mis en cascade. La valeur finale affichée correspond à la valeur médiane entre le dernier seuil activité et le seuil directement supérieur. Si le premier modèle affiche une prédiction de type Basse, alors la sortie sera la moyenne de la saison considérée. Par exemple, dans le problème de détermination de la turbidité à l'eau brute pour la station

Des Baillets, si les modèles classifiant de la saison en cours donnent les résultats « Haute/Haute/Basse/Basse », respectivement pour les seuils de 4/5,5/7,5/9,3 UTN, le résultat de synthèse affiché pour la classification sera 6,5 UTN. Deuxième exemple, si le résultat est « Basse/Basse/Haute/Haute », et que la saison est l'automne, alors la sortie sera la moyenne de la turbidité à Des Baillets à l'automne, soit 1,97 UTN. Une illustration pour la turbidité à l'eau brute de Des Baillets figure au Tableau 3-15.

Regroupement des modèles saisonniers – prédiction tout au long de l'année

À partir de ce point, trois modèles saisonniers ont été créés, il faut les regrouper pour fournir une prédiction tout au long de l'année. La date de transition d'une saison à l'autre n'est pas fixe : effectivement, d'une année à l'autre l'été prend fin plus ou moins tôt, la fonte des neiges est plus ou moins tardive, etc. Ces périodes de transitions floues sont surtout problématiques pour les passages des évènements de faible turbidité à des saisons riches en pointes : i.e., fin automne – printemps (janvier - février), et été – début automne (septembre - octobre).

Afin de représenter le flou que l'on peut avoir sur ces transitions, et dans le but d'agir de manière conservatrice dans la prédiction fournie à l'opérateur, celui-ci sera informé chaque jour de la prédiction fournie par chaque modèle saisonnier, soit trois prédictions. La décision d'accorder plus ou moins d'importance à la prédiction d'une saison plutôt qu'une autre, sera orientée par une probabilité de pertinence des modèles saisonniers. Entre les dates fixes identifiées lors de la phase d'analyse et de découpage temporel, la pertinence du modèle adéquat sera de 100% (exemple : modèle été avec une pertinence de 100% en juillet). Au voisinage de ces dates de transition, la probabilité du modèle en fin de saison baissera en même temps que celle du modèle de la saison à venir augmentera, telle qu'illustré à la Figure 2-7.

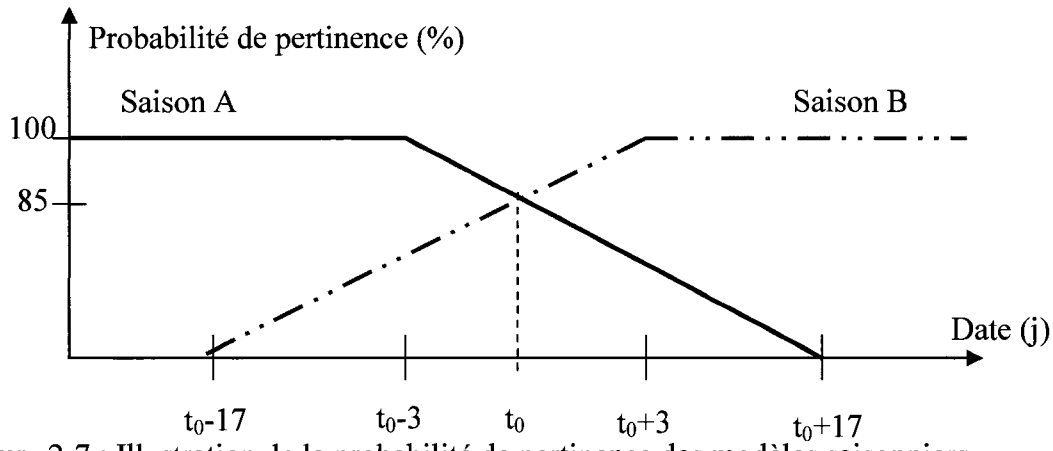

Figure 2-7 : Illustration de la probabilité de pertinence des modèles saisonniers

La date t_0 représente la date fixe identifiée comme étant la transition entre deux périodes. Les valeurs suivantes ne sont pas issues de mesures, elles ont été choisies arbitrairement par le modélisateur. La probabilité du modèle en fin de saison (modèle A) décroît de 5% par jour pendant 20 jours. Elle commence à décroître trois jours avant t_0, et atteint 0% de pertinence 17 jours après. La saison qui débute (ici, saison B) agit de manière symétrique par rapport à t_0. Ces dates arbitraires ont été choisies de sortes que les modèles soient à 50% dans les 10 jours autour de t_0. Ceci semble un compromis raisonnable entre apporter du flou relativement à l'échelle des données observées à l'eau brute en fonction du temps sur dix années (Figure 3-1), et fournir des modèles spécifiques dont l'interprétation des prédictions ne sera pas ambigüe pour l'opérateur.

De plus, pour des périodes sensibles de transition, i.e. le passage d'une saison faible en turbidité à une saison riche en pointes ou vice versa, des règles spéciales pourront mettre l'accent sur le modèle spécialisé sur les pointes (afin de rester conservateur).

La décision finale dans la zone de flou fera appel au bon sens de l'utilisateur : si une année exceptionnellement froide et abondante en chutes de neige ne voit ses températures monter au-dessus de 0 °C qu'à partir de la mi-mai (au lieu d'avril), la fonte des neiges se trouve retardée. Ainsi, le modèle printanier (fortement dépendant

de cette fonte) sera à prendre en considération, même si la date favoriserait le modèle été.

En conclusion de ce chapitre, les méthodes utilisées dans ce projet pour élaborer un modèle par réseau de neurones ont été définies points par points. Les grandes étapes de la démarche consistent à définir dans un premier temps les besoins auxquels devra répondre le modèle, et en fonction de ces derniers, le type de modèle le plus adapté sera choisi. Ensuite, les connaissances préalable sur le sujet (ou la revue de littérature) permettent de réunir une base de données préliminaire contenant toutes les entrées potentielles du ou des modèles. Une fois cette base de données nettoyée des entrées ou exemples non pertinents, ces derniers seront répartis en deux groupes : les exemples permettant la calibration des modèles, et les exemples non présentés au réseau lors de la phase de calibration (i.e. permettant de tester la capacité de généralisation du modèle). Puis, des méthodes statistiques permettent de sélectionner un sous-ensemble d'entrées contenant le plus d'information pour la prédiction. Plusieurs réseaux sont ensuite calibrés avec ces entrées, ils sont comparés entre eux sur la base d'un critère de performance adapté aux besoins du problème. Finalement, les modèles finaux retenus sont mis en commun afin de fournir une prédiction tout au long de l'année.

Chapitre 3 PRÉDICTION DE LA TURBIDITÉ À LA PRISE D'EAU BRUTE DE LA STATION CHARLES J. DES BAILLETS

Après avoir vu les éléments constitutifs de la méthodologie adoptée, ce chapitre rappelle brièvement les choix effectués et présente les résultats des modèles pour l'usine Charles J. Des Baillets.

3.1 Étapes de la modélisation

L'ensemble des données récupérées est décrit au Tableau 3-1. Sauf indication contraire, celles-ci sont disponibles du 1er janvier 1996 au 31 mai 2006 et correspondent à des moyennes journalières.

Les codes utilisés pour décrire les variables sont entre parenthèses. Pour alléger le tableau, tous les codes des paramètres météorologiques ne sont pas répétés, car ils suivent tous le même schéma : « Lieu_Type de variable mesurée». La nomenclature pour décrire un lieu ou un type de variable mesurée n'est précisée que la première fois où elles apparaissent.

Tableau 3-1 : Résumé des 40 variables disponibles initialement et des codes utilisés pour décrire ces variables

Qualité de l'eau	Données hydrologiques	Paramètres météorologiques
A MONTRÉAL (ATW ET DB) : • turbidité EB *(TURB_DB)* • couleur EB *(COUL_DB)* • température EB *(TEMP_DB)* • conductivité ET *(COND_DB)* A HAWKESBURY : • turbidité EB *(TURB_HAW)* • couleur EB *(COUL_HAW)* • température EB *(TEMP_HAW)* A OKA (1er janvier 1998 au 31 mai 2006) : • température EF *(TEMF_OKA)* A BEAUHARNOIS (1er janvier 2001 au 31 décembre 2005) : • turbidité EB *(TURB_BHN)*	DÉBIT DES OUTAOUAIS à : • Vaudreuil *(OUT_VAUD)* • Ste Anne de Bellevue *(OUT_SAB)* • Barrage de Carillon *(OUT_CARI)* DÉBIT DU FLEUVE ST-LAURENT à : • Lasalle *(FLV_LSL)* • Des Cèdres *(FLV_CED)* • Barrage de Beauharnois *(FLV_BHN)* DÉBIT DE TRIBUTAIRES SECONDAIRES : • rivière des Raisins à Glen Navis *(RIV_RAIS)* • rivière Beaudette à Williamstown *(RIV_BEAU)* • rivière Châteauguay à Châteauguay *(RIV_CHAT)*	VITESSE DU VENT, moyenne horaire et maximale horaire : • à Dorval *(DOR_VITM et DOR_VITX)* • au lac St François *(LSF_...)* • à Ste Anne de Bellevue *(SAB_...)* PRÉCIPITATION MOYENNE JOURNALIÈRE : • à Dorval *(DOR_PREC)* • au lac St François • à Ste Anne de Bellevue • à Rigaud *(RIG_...)* • à Valleyfield *(VAL_...)* • aux Cèdres *(CED_...)* TEMPÉRATURE MOYENNE JOURNALIÈRE DE L'AIR : • à Dorval *(DOR_TEMP)* • au lac St François • à Ste Anne de Bellevue • à Rigaud • à Valleyfield • aux Cèdres

3.1.1 Définition d'un évènement turbide

Conformément à la méthodologie développée par Tremblay (2004), les données de turbidité furent discrétisées à partir d'une analyse des quantiles (séparation en cinq classes de turbidité). Ainsi la notion d'évènement « *turbide* » est définie comme étant les exemples dont la turbidité moyenne journalière à l'eau brute de la station Des

Baillets dépasse 3,1 UTN, soit le 90^{e} centile ou les classes III, IV et V. Les résultats de cette analyse figurent au Tableau 3-2.

Tableau 3-2 : Classes de turbidité, eau brute de la station Des Baillets

Classe de turbidité	Centiles	Bornes de turbidité de l'intervalle (UTN)
I	0 à 0,75	TURB_DB < 2,3
II	0,75 à 0,90	2,3 ≤ TURB_DB < 3,1
III	0,90 à 0,95	3,1 ≤ TURB_DB < 4
IV	0,95 à 0,99	4 ≤ TURB_DB < 9,3
V	Entre 0,99 et 1	9,3 ≤ TURB_DB < 31,3

Notons que pour l'application pratique, le saut de 4 UTN à 9,3 UTN est assez brusque. Une classification par seuil sera utilisée, mais en définissant des valeurs intermédiaires : 4 UTN, 5,5 UTN, 7,5 UTN et 9,3 UTN.

Une distinction temporelle est aussi émise : un évènement turbide est dit « *de fond* » si sa durée excède ou est égale à 5 jours. Sinon, il est appelé de « *courte durée* ».

Un évènement turbide de courte durée seul sera dit « *isolé* ». Cependant, s'il arrive en même temps qu'un évènement turbide de fond, il est appelé « *superposé* » (Figure 3-2).

3.1.2 Analyse préliminaire de la variables de sortie : TURB_DB

Le tracé de la turbidité en fonction du temps pour dix années de données à la Figure 3-1 fait ressortir l'existence de cinq « saisons » où la réponse du système semble donner des turbidités à l'eau brute différentes. Aucune périodicité ne se démarque clairement.

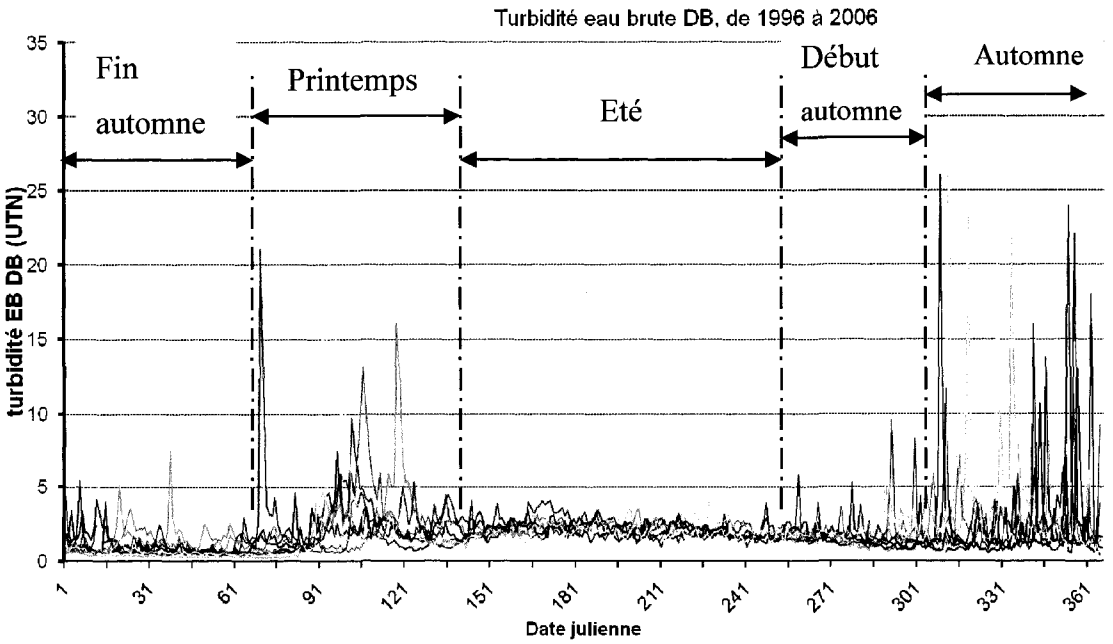

Figure 3-1 : Turbidité de l'eau brute à Des Baillets de 1996 à 2006, en fonction de la date julienne

Trois grandes périodes sont observées (automne, printemps et été) avec deux périodes de transition autour de l'automne. La première appelée « début automne » résulte en une dégradation progressive de la qualité de l'eau au sortir de l'été, sans doute sous l'influence des premières tempêtes. La seconde, « fin automne », serait probablement liée à des évènements de type automne en l'absence ou pendant la fragilisation du couvert de glace (protection aux intempéries). Ceci expliquerait les pointes de janvier février 2006, hiver chaud et pluvieux où la saison d'automne s'est prolongée pendant longtemps. La localisation temporelle de ces périodes de transitions varie d'une année à l'autre. Un détail du printemps et de l'automne (et ses périodes de transition) est donné à la Figure 3-2 et à la Figure 3-3. Les caractéristiques générales des saisons ainsi que les dates de découpage graphique figurent dans le Tableau 3-3. Les statistiques descriptives sont données en Tableau 3-4.

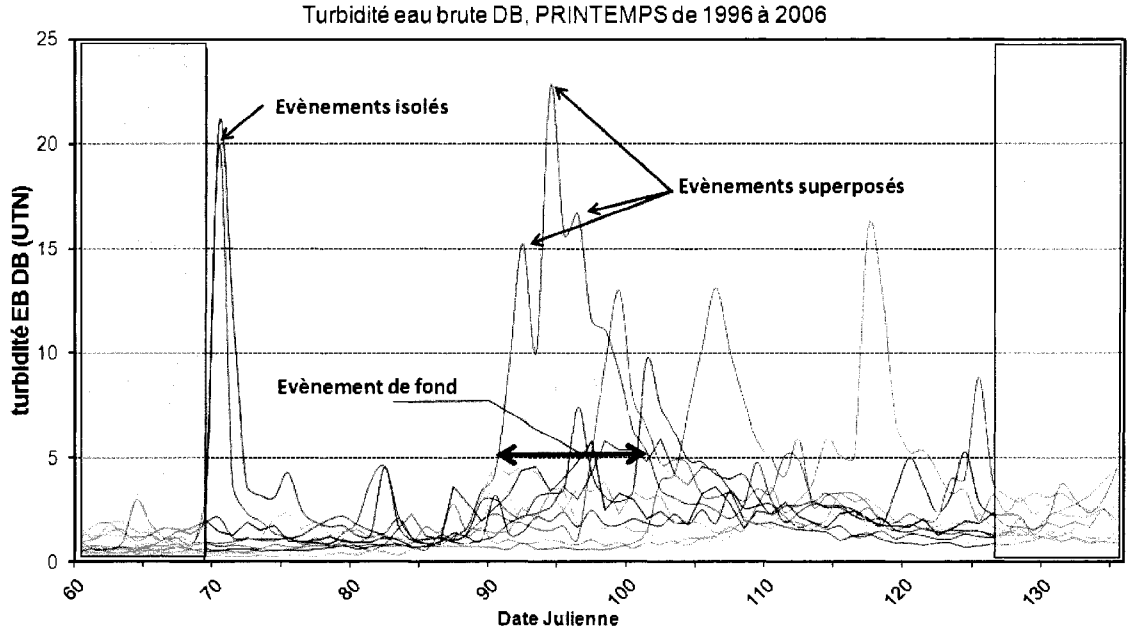

Figure 3-2 : Turbidité de l'eau brute à Des Baillets au printemps, de 1996 à 2006, en fonction de la date julienne

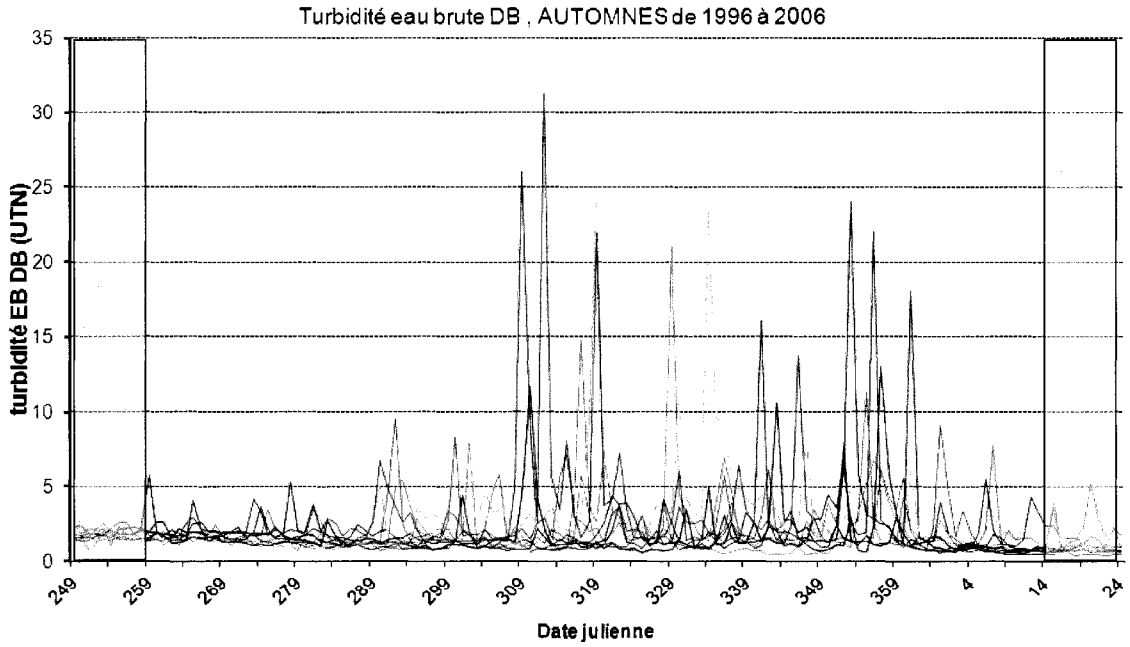

Figure 3-3 : Turbidité de l'eau brute à Des Baillets à l'automne et ses périodes de transition, de 1996 à 2006, en fonction de la date julienne

Tableau 3-3 : Découpage graphique en cinq saisons

Nom de la période	Caractéristiques générales	Dates
Fin automne	Période de transition. Quelques pointes d'amplitude moyenne (max 7,8 UTN).	1er janv. – 9 mars
Printemps	Fortes turbidités. Pointes isolés, superposés et de fond.	10 mars – 31 mai
Été	Turbidité moyenne faible mais plus élevée qu'à l'hiver. Quelques pointes de turbidité de faible amplitude (< 5 UTN).	1er juin – 15 sept.
Début automne	Période de transition. Quelques pointes d'amplitude moyenne (max 9,6 UTN).	16 sept. – 31 oct.
Automne	Nombreuses pointes de turbidité de grande amplitude. Augmentations soudaines et de courte durée. Majoritairement des évènements isolés, ponctué de quelques évènements de fond.	1er nov. – 31 déc.

Tableau 3-4 : Statistiques descriptives de TURB_DB

	TURB_DB (en UTN)									
	N	Moyenne	Médiane	Minimum	Maximum	P25	P75	P90	P95	Ecart type
Toutes les données	3770	1,96	1,67	0,27	31,30	1,08	2,30	3,03	4,10	1,86
Printemps	616	2,52	1,81	0,27	22,80	1,08	3,00	4,80	7,40	2,62
Eté	1334	2,09	2,00	0,81	4,90	1,70	2,40	2,80	3,10	0,58
Automne	1323	1,97	1,42	0,37	31,30	1,04	2,00	3,30	4,70	2,36

3.1.3 Tri de la base de données

Le tri de la base de données permet de retenir les 27 variables candidates pour l'année entière : respectivement 6, 9 et 12 variables pour les catégories qualité de l'eau, débit, et météorologie. À cause du manque de données, la qualité de l'eau à Beauharnois, et les variables météorologiques à Des Cèdres, Rigaud, et Valleyfield ont été écartées. De plus, les mesures de la température de l'air en trois places différentes furent

regroupées en une seule issue de la moyenne arithmétique : la température de l'air étant relativement constante sur la région d'étude à une date donnée.

3.1.4 Choix des variables d'entrées

Connaissances préalables des causes de la turbidité

Basé sur les résultats de Tremblay (2004), huit facteurs explicatifs des évènements turbides furent identifiés, et trois variables pour représenter trois de ces facteurs furent créées (voir l'Annexe A pour les détails de construction de ces index). Ces index ont été légèrement modifiés par rapport aux travaux antérieurs. Ces causes explicatives ainsi que les valeurs seuils d'activation pour l'observation graphique sont rappelées dans le Tableau 3-5. Ces valeurs seuils furent pour la plupart identifiées par Tremblay (2004), elles correspondent généralement à une valeur supérieure à la médiane ou au 85^{e} centile.

Tableau 3-5 : Causes, variables explicatives, et seuils d'activation pour l'analyse graphique

Cause	Variables explicatives	Valeur seuil d'activation de ces variables
Dégradation de la qualité de l'eau en amont	TURB_DB TURB_HAW	Une pointe de turbidité (supérieure au 90e centile) est observée les jours précédents à Des Baillets et Hawkesbury. TURB_DB ≥ 3,1 UTN TURB_HAW ≥ 9.7 UTN
Fonte des neiges (printemps)	IDX_FONT	= 1
Fragilité du couvert de glace (printemps / fin automne)	IDX_FONT	= ½
Renversement (printemps et automne)	IDX_RENV	situé entre 0,3 et 1
Pluie	DOR_PREC SAB_PREC LSF_PREC	Pluie ≥ 5 mm
Vent	(DOR, SAB, et LSF)_ VITM et VITX	Vent moyen (_VITM) ≥ 25 km/h Vent maxi (_VITX) ≥ 35 km/h
Hausse des tributaires secondaires	RIV_(BEAU, RAIS, et CHAT)	RIV_BEAU ≥ 5 m^3/d RV_RAIS ≥ 9 m^3/d RIV_CHAT ≥ 50 m^3/d
Contribution des Outaouais	OUT_FLV	OUT_FLV ≥ 6 % (médiane sur l'année)

Recensement des évènements turbides

Sur nos 3827 exemples disponibles à Des Baillets, nous avons recensé 213 cas de turbidité (TURB_DB ≥ 3,1 UTN, classes III, IV ou V). Ces cas sont répartis comme suit (Tableau 3-6) :

- Les années 1998, 2002, et 2003 eurent le plus de pointes.

- Le début de l'automne et la fin de l'automne sont bien des périodes de transition de l'automne. La plupart des évènements turbides sont isolés.
- Les trois principales saisons sont l'automne, l'été et le printemps.
- L'automne et le printemps cumulent une dizaine d'évènements de fond, presque autant que d'années, il doit s'agir d'évènements se produisant une fois par an (renversement, etc.).
- L'automne se distingue par une abondance d'évènements isolés alors que le printemps est plutôt constitué d'évènements de fond et superposés.

Tableau 3-6 : Recensement des évènements turbides (≥3,1 UTN) par saison de 1995 à 2006

		1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	Totaux	Total par saison
Fin automne	**F**				1									1	12
	S				1									1	
	I	1		2			2		2		1		2	10	
Printemps	**F**		3	2	1	1	2	1	1	1	1	1		14	80
	S		4	8	3	3	4	2	2	3	3	4		36	
	I		1	2	3	2		1	7		3	3	8	30	
Été	**F**				1				1					2	36
	S				2				5					7	
	I		3	1	5	3	2	1	3	7	2			27	
Début automne	**F**													0	19
	S													0	
	I			1	2	1	2	2	3	4		4		19	
Automne	**F**									3		2		5	66
	S									5		4		9	
	I		7	2	7	8	5	4	5	7	5	2		52	
Total par année		1	18	18	26	18	17	11	29	30	15	20	10		213

* F = Fond; S = Superposé; et I = Isolé

En Annexe B figurent les tableaux du nombre et le pourcentage d'occurrences de nos facteurs explicatifs. La cause « qualité de l'eau en amont » fut omise, car elle apparaît bien souvent pour les évènements de fond (par définition même). Les facteurs dont l'occurrence dépasse 51% sont indiqués en gras. Il est important de rappeler que des évènements n'ayant lieu qu'une seule fois par saison comme le renversement ou la fonte printanière ne peuvent avoir un pourcentage d'occurrences élevé, il faudra alors regarder le nombre d'occurrences par rapport au nombre de saisons d'observations. Plusieurs commentaires résultent de ces tableaux :

- Durant la fin de l'automne, seulement un évènement de fond et un superposé sont recensés. Cependant pour les évènements isolés, la contribution des Outaouais et le vent sont des facteurs importants. Il serait possible que lors de fortes pluies et de la fragilisation du couvert de glace, la protection offerte contre le vent s'amenuiserait en même temps que la hausse des débits et la régulation forcée de la contribution des Outaouais par le barrage de Carillon dégraderait la qualité générale de l'eau. Il pourrait s'agir d'années exceptionnellement chaudes ou les causes automnales durent plus longtemps.
- Durant le printemps, la fonte des neiges est accompagnée de la hausse des contributions des tributaires et de la contribution des Outaouais qui restent élevés toute la saison. Sur les dix années d'étude, sept années font coïncider un évènement de fond avec le renversement et la fonte des neiges; alors que l'indicateur fournit par la contribution des Outaouais est actif douze fois sur les quatorze évènements de fond recensés. Précipitations et tempêtes de vent semblent avoir aussi un rôle à jouer dans les pointes isolées de turbidité.
- Durant l'été, les causes majoritaires semblent liées aux précipitations et à la hausse résultante des tributaires secondaires.
- Durant le début de l'automne, seuls des évènements isolés sont recensés. Ils sont associés à des tempêtes de vent et des précipitations. Cette saison semble être fortement liée à l'automne.

- Durant l'automne, (ainsi que début et fin), les évènements isolés ont pour causes majeures les tempêtes de vent, puis les précipitations accompagnées d'une hausse des tributaires, et la contribution des Outaouais dans une moindre mesure. En ce qui concerne le renversement, sa contribution à la dégradation de la qualité moyenne de l'eau semble minime vis-à-vis des pointes isolées de forte amplitude.

Ces observations confirment une dégradation générale de la qualité de l'eau pour les périodes d'automne, du printemps, et dans une moindre mesure, de l'été. Il semble que les évènements de fond soient liés à des indicateurs comme le renversement et la contribution de la rivière des Outaouais (pour l'automne et le printemps). À cette dégradation de la qualité moyenne viennent se superposer des évènements ponctuels de forte amplitude (vent avec ou sans pluie). Ces observations soutiennent l'élaboration de modèles distincts de classification à des seuils de turbidité croissants. Chaque modèle serait donc spécifique à un seuil et pourrait distinguer les causes propres à la turbidité pour diverses amplitudes.

Choix des décalages temporaux

Aux 26 variables candidates viennent s'ajouter possiblement leurs décalages temporaux. Les temps de réponse du système étant assez rapide, seuls les décalages de un à x jours dans le passé seront considérés. Pour les données de débit et de qualité de l'eau en amont, l'eau aurait un temps de séjour de deux à trois jours entre Cornwall et Portneuf (Couillard, 1987); ainsi, une valeur double de six jours pour la fenêtre temporelle semble conservatrice. Tremblay avait retenu dans ses modèles l'influence du vent décalé de un jour (Tremblay, 2004). Une fenêtre de trois jours pour le vent sera envisagée. L'examen graphique de la turbidité en fonction du temps (voir la section sur le choix des variables d'entrées), donne les mêmes résultats et a révélé l'influence potentielle de la pluie, de l'index de renversement et de l'index de fonte (fragilité ou bris du couvert de glace) jusqu'à 6, 10, et 17 jours respectivement. Notons

que la valeur 17 jours est unique car la plupart des décalages observés pour la fonte se situent entre un et dix jours avant l'évènement turbide.

Les fenêtres temporelles considérées, pour former la base de données préliminaire, sont récapitulées entre parenthèses dans le Tableau 3-7. Ceci représente un total de 91 variables potentielles d'entrée.

Tableau 3-7 : Fenêtre temporelle des variables de la base de données préliminaire

Qualité de l'eau **4 variables**	**Débit** **10 variables**	**Météorologie** **10 variables**	**Index** **2 variables**
COUL_DB (3) TURB_DB (3) COND_DB (3) TURB_HAW (3)	OUT_CARI (3) OUT_VAUD (3) OUT_SAB (3) FLV_LSL (3) FLV_BEAU (3) FLV_CED (3) OUT_FLV (3) RIV_RAIS (7) RIV_BEAU (7) RIV_CHAT (5)	DOR_VITM (3) DOR_VITX (3) LSF_VITM (3) LSF_VITX (3) SAB_VITM (3) SAB_VITX (3) DOR_PREC (6) LSF_PREC (6) SAB_PREC (6) PRECX_DS (6)	IDX_RENV (10) IDX_FONT (10)

Nouveau découpage temporel

Dans un premier temps, une analyse discriminante est menée sur l'année entière avec toutes les variables de la base de données préliminaire plus une variable texte : saison. Dépendamment du seuil considéré, cette variable de saison ressort souvent dans les dix premiers descripteurs (sur la vingtaine que vont retenir les modèles). Par conséquent, le découpage en saison est une bonne idée. Avec les connaissances acquises, un deuxième découpage en saison plus précis va être réalisé a posteriori, afin d'aider le réseau dans la reconnaissance d'exemples choisis.

Étant donné que les trois sous-saisons d'automne semblent présenter des caractéristiques et des causes explicatives très proches, et pour ne pas trop réduire le

nombre d'exemples disponibles par saison (et altérer la capacité d'apprentissage du modèle), elles sont regroupées.

Le tracé du nuage de points des données de turbidité en fonction de la date en journée julienne donne un découpage plus précis que celui opéré précédemment. Cette figure n'apporte pas plus d'information utile à la compréhension du phénomène, elle ne sera pas présentée ici. Seules les trois saisons majeures seront considérées (printemps, été, automne) et quatre modèles par saison seront bâtis (un par valeur seuil de turbidité). Ce deuxième découpage saisonnier est indiqué dans le Tableau 3-8. Ce nouveau découpage sera considéré pour le reste de l'étude.

Les frontières avec l'été ont été choisies de sorte que les évènements estivaux soient inférieurs à 5 UTN (TURB_DB). L'été n'étant pas une saison possédant des pointes de turbidité abruptes et de forte amplitude, nous ne développerons pas de modèle de classification spécialisé pour cette saison : un modèle de régression étant suffisant. Nos nouvelles dates sont très proches de celles identifiées précédemment.

La saison dite « hiver » utilisera le modèle « automne » car il s'agit d'évènements automnaux durant les années chaudes (sans la protection offerte par le couvert de glace). Les exemples de l'hiver n'ont pas été inclus dans la conception du modèle automne afin de spécialiser celui-ci dans l'apprentissage des pointes de la saison.

Tableau 3-8 : Deuxième découpage en saison

Saison	Date de début	Date de fin
Printemps	11 mars	5 mai
Été	6 mai	15 septembre
Automne	16 septembre	24 janvier
Hiver	25 janvier	10 mars

Analyse discriminante par saison

L'analyse discriminante est menée par saison et par seuil de turbidité. Chaque analyse retient un sous-ensemble de variables d'entrée étant les meilleurs descripteurs pour séparer linéairement les classes de turbidité basse et haute. Les variables retenues sont indiquées dans les trois tableaux suivants, un décalage temporel étant représenté par la notation « $VARIABLE_i$-x » symbolisant la variable i à la date t-x jours. Ainsi, TURB_DB-1 représente la turbidité de l'eau brute à Des Baillets la veille.

Certaines variables semblent influentes pour tous les seuils (par exemple, LSF_VITM-1 à l'automne) alors que d'autres sont spécifiques à certains seuils (PRECX_DS-1 et -5 pour les pointes supérieurs à 7,5 UTN à l'automne).

Avis complémentaire – analyse par réseau de neurones de type PNN

Les résultats des analyses discriminantes, « *Backward Stepwise* » et « *GAPNN* » sont fournis, pour les saisons automne et printemps, dans les tableaux récapitulatifs ci-dessus. En Annexe B, figure un tableau récapitulatif des commentaires émis lors de l'interprétation des entrées par analyse avec réseaux de type PNN (Tableau A-3).

À titre informatif, les variations observées sur le critère de performance en fonction des entrées sélectionnées sont de l'ordre de 1%. Il n'y a ainsi pas de distinction (ou de saut) notable sur l'information apportée par telle ou telle variable. En effet, nous obtiendrons de bonnes performances avec des variables qui furent éliminées par l'analyse « *Stepwise Backward* ».

Tableau 3-9 : Entrées sélectionnées par analyse statistique – automne

AUTOMNE									
Liste des variables	Seuil de turbidité (en UTN)				Liste des variables	Seuil de turbidité (en UTN)			
	4	5,5	7,5	9,3		4	5,5	7,5	9,3
TURB_DB-1	DB	D			DOR_PREC-1				
TURB_DB-2					DOR_PREC-2		D		
TURB_DB-3		D	D		DOR_PREC-3				
COUL_DB-1					DOR_PREC-4				
TURB_HAW-1	D	DG	D	D	DOR_PREC-5				
TURB_HAW-2		D	D	D	DOR_PREC-6				
TURB_HAW-3					LSF_PREC-1				
OUT_FLV-1					LSF_PREC-2		D		
OUT_FLV-2					LSF_PREC-3				
RIV_RAIS-1	D	D			SAB_PREC-1		D		
RIV_RAIS-2	D			D	SAB_PREC-2				
RIV_RAIS-3		D			PRECX_DS-1			D	D
RIV_RAIS-4					PRECX_DS-2				
RIV_RAIS-5					PRECX_DS-3				
RIV_RAIS-6					PRECX_DS-4				
RIV_RAIS-7					PRECX_DS-5			D	D
RIV_BAUD-1				D	PRECX_DS-6				
RIV_BAUD-2					IDX_RENV-1				D
RIV_BAUD-3					IDX_RENV-2				
RIV_BAUD-4					IDX_RENV-3				
RIV_BAUD-5					IDX_RENV-4				
RIV_BAUD-6					IDX_RENV-5				
RIV_CHAT-1				D	IDX_RENV-6				
RIV_CHAT-2					IDX_FONT-1	D			D
RIV_CHAT-3					IDX_FONT-2				
RIV_CHAT-4					IDX_FONT-3	D			D
RIV_CHAT-5					IDX_FONT-4				
DOR_VITM-1									
DOR_VITM-2	BG	BG							
DOR_VITX-1									
DOR_VITX-2	BG	BG	DBG	B					
LSF_VITM-1	DBG	BG	DBG	BG					
LSF_VITM-2									
LSF_VITX-1		D							
LSF_VITX-2	B	BG	DBG	D					
SAB_VITM-1	BG	DBG	BG	DBG					
SAB_VITM-2	B	G	D	D					
SAB_VITX-1	D								
SAB_VITX-2									

Légende

D : analyse discriminante
B : "Stepwise Backward - PNN"
G : "Genetic Algorithm - PNN"

Tableau 3-10 : Entrées sélectionnées par analyse statistique - printemps

Printemps									
Liste des variables	Seuil de turbidité (en UTN)				Liste des variables	Seuil de turbidité (en UTN)			
	4	5,5	7,5	9,3		4	5,5	7,5	9,3
TURB_DB-1	DBG	DBG	DBG	DBG	DOR_PREC-1	D			
TURB_DB-2				D	DOR_PREC-2	D	D	D	D
TURB_DB-3	DBG		G		DOR_PREC-3				
COUL_DB-1					DOR_PREC-4				
TURB_HAW-1					DOR_PREC-5				
TURB_HAW-2					DOR_PREC-6				
TURB_HAW-3	DBG	DBG	B		LSF_PREC-1				
OUT_FLV-1		B	DBG	DBG	LSF_PREC-2	D	D	D	G
OUT_FLV-2					LSF_PREC-3		D		
RIV_RAIS-1			D		SAB_PREC-1		D		
RIV_RAIS-2					SAB_PREC-2				
RIV_RAIS-3			D		PRECX_DS-1	G			
RIV_RAIS-4					PRECX_DS-2				
RIV_RAIS-5					PRECX_DS-3		D		
RIV_RAIS-6	DBG	BG	BG	BG	PRECX_DS-4				
RIV_RAIS-7					PRECX_DS-5				
RIV_BAUD-1					PRECX_DS-6		D		D
RIV_BAUD-2					IDX_RENV-1	DG	B	B	B
RIV_BAUD-3			D		IDX_RENV-2				
RIV_BAUD-4					IDX_RENV-3				
RIV_BAUD-5					IDX_RENV-4				
RIV_BAUD-6					IDX_RENV-5			D	
RIV_CHAT-1	DBG		DG		IDX_RENV-6				
RIV_CHAT-2					IDX_FONT-1	DBG	BG	BG	BG
RIV_CHAT-3	G	DBG	DG	DBG	IDX_FONT-2	BG	BG	DBG	G
RIV_CHAT-4					IDX_FONT-3	BG	BG	BG	D
RIV_CHAT-5	D	BG		DBG	IDX_FONT-4	DBG	G	D	
DOR_VITM-1									
DOR_VITM-2			D	D					
DOR_VITX-1									
DOR_VITX-2									
LSF_VITM-1	DB	DB	DB	DG					
LSF_VITM-2									
LSF_VITX-1									
LSF_VITX-2			D						
SAB_VITM-1	B		B	G					
SAB_VITM-2									
SAB_VITX-1									
SAB_VITX-2									

Légende

D : analyse discriminante
B : "Stepwise Backward - PNN"
G : "Genetic Algorithm - PNN"

Tableau 3-11 : Entrées sélectionnées par analyse statistique - été

Eté					
Liste des variables	Seuil de turbidité (en UTN)		Liste des variables	Seuil de turbidité (en UTN)	
	4	Reg		4	Reg
TURB_DB-1	B	BG	DOR_PREC-1		
TURB_DB-2			DOR_PREC-2		
TURB_DB-3		BG	DOR_PREC-3		
COUL_DB-1	D		DOR_PREC-4		
TURB_HAW-1			DOR_PREC-5		
TURB_HAW-2			DOR_PREC-6		
TURB_HAW-3			LSF_PREC-1		
OUT_FLV-1			LSF_PREC-2	DG	
OUT_FLV-2			LSF_PREC-3		
RIV_RAIS-1			SAB_PREC-1		
RIV_RAIS-2			SAB_PREC-2		
RIV_RAIS-3			PRECX_DS-1		BG
RIV_RAIS-4			PRECX_DS-2		
RIV_RAIS-5			PRECX_DS-3		
RIV_RAIS-6			PRECX_DS-4		
RIV_RAIS-7			PRECX_DS-5		
RIV_BAUD-1	D		PRECX_DS-6		
RIV_BAUD-2			IDX_RENV-1		
RIV_BAUD-3			IDX_RENV-2		
RIV_BAUD-4			IDX_RENV-3		
RIV_BAUD-5			IDX_RENV-4		
RIV_BAUD-6			IDX_RENV-5		
RIV_CHAT-1	DB	BG	IDX_RENV-6		
RIV_CHAT-2	D		IDX_FONT-1		
RIV_CHAT-3	B	B	IDX_FONT-2		
RIV_CHAT-4			IDX_FONT-3		
RIV_CHAT-5			IDX_FONT-4		
DOR_VITM-1					
DOR_VITM-2					
DOR_VITX-1					
DOR_VITX-2					
LSF_VITM-1	D				
LSF_VITM-2					
LSF_VITX-1					
LSF_VITX-2					
SAB_VITM-1					
SAB_VITM-2					
SAB_VITX-1					
SAB_VITX-2					

Légende

D : analyse discriminante
B : "Stepwise Backward - PNN"
G : "Genetic Algorithm - PNN"

Choix final des entrées des modèles

Pour chaque saison et pour chaque seuil, la sélection donnée par l'analyse discriminante précédente est retenue, accompagnée des commentaires ci-dessous :

- Plus de données sont disponibles pour RIV_RAIS que pour RIV_BEAU, ces deux rivières étant proches géographiquement, l'utilisation de RIV_RAIS avec le même décalage temporel que celui préconisé pour RIV_BEAU sera privilégiée.
- Les données de qualité de l'eau et de débit sont fortement corrélées d'un jour au suivant. Pour éviter la redondance de l'information, nous veillerons à laisser une journée d'écart entre deux mesures. Exemple : le modèle automne préconise RIV_RAIS-1, -2 et -3 en deçà de 5,5 UTN. Seulement les décalages -1 et -3 seront conservés, l'information détenue par -2 étant redondante. Les données de météorologie sont beaucoup moins auto-corrélées, cette restriction ne s'applique donc pas pour elles.

3.1.5 Partitionnement des exemples

Les détails du partitionnement sont fournis à l'Annexe C. Il s'agit de la nomenclature adoptée, du nombre d'exemples disponibles pour chaque saison et chaque ensemble, des années de test et de sélection pour les échantillonnages à année fixe, et des proportions de répartition des exemples pour les échantillonnages aléatoires.

3.1.6 Choix d'une architecture de réseau, des paramètres internes, et calibration des réseaux

Se référer au tableau récapitulatif de l'Annexe F.

3.1.7 Choix d'un critère de performance

La définition des critères de performance créés pour les besoins du problème, ainsi qu'un tableau récapitulatif des valeurs employées figurent à l'Annexe G.

3.1.8 Détermination du meilleur réseau

Exemple de sélection d'une liste de nombre de neurones cachés optimaux

Voici l'exemple de la prédiction du seuil 9.3 UTN au printemps.

L'analyse des entrées sans prétraitement ayant déjà été effectuée, après quelques essais et erreurs, nous avons retenu l'ensemble d'entrées 031 (10 entrées : TURB_DB-1 et -2; OUT_FLV-1; RIV_CHAT-3 et -5; DOR_VITM-2 et LSF_VITM-1; DOR_PREC-2 et PRECX_DS-6; IDX_FONT-3). Nous désirons évaluer le même groupe, mais prétraité 031* (* signifiant « avec prétraitement »). Nous commencerons à l'étape 4 de notre méthodologie (voir la section 2.3.10). Deux répartitions sont considérées ainsi que quatre critères de performance. Deux listes de nombre de neurones donnant de bons résultats sont retenues.

Nous avons tracé les diagrammes boîtes à moustaches pour les quatre critères et les les boîtes de neurones les plus performants sont surlignées.

En Figure 3-4, pour la répartition 021, le minimum des pertes globales (fig. b) est obtenu pour la liste {7; 9; 11; 18; 26; 32; 44; 66}, avec moins de variations pour {11; 18; 26; 66}. Les pertes sur TEST proposent la liste {1; 5; 6; 8; 14; 22}. Choisir peu de neurones (notamment les solutions 5 ou 6 neurones), donne une bonne généralisation (fig. a), mais l'apprentissage est sans doute incomplet car pas assez de complexité n'est apportée au modèle en deçà de 7 neurones (les pertes sommées décroissent jusqu'à 7 neurones en Figure b). Toutefois, certaines solutions performantes peuvent être obtenues occasionnellement pour 5 ou 6 neurones. En ce qui concerne la

maximisation du critère matrice de performance TEST, celui-ci limite nos choix à {1; 8; 22; 62}. Le pourcentage de classification correcte TEST, semble privilégier une zone moins étendue. {7; 9; 10; 11 16; 32} neurones se démarquent, bien que toutes les performances soient assez bonnes : i.e., supérieure à 70%. Nous retiendrons pour cette répartition le plus petit ensemble faisant l'unanimité, soit entre 7 et 9 neurones.

Pour la répartition 001, les résultats sont très similaires bien que 9 neurones se démarquent plus nettement.

En comparant les performances moyennes de l'effet du prétraitement sur nos deux répartitions, nous observons que les résultats vont dans le même sens; ainsi, seul la répartition 021 sera commentée en Figure 3-5.

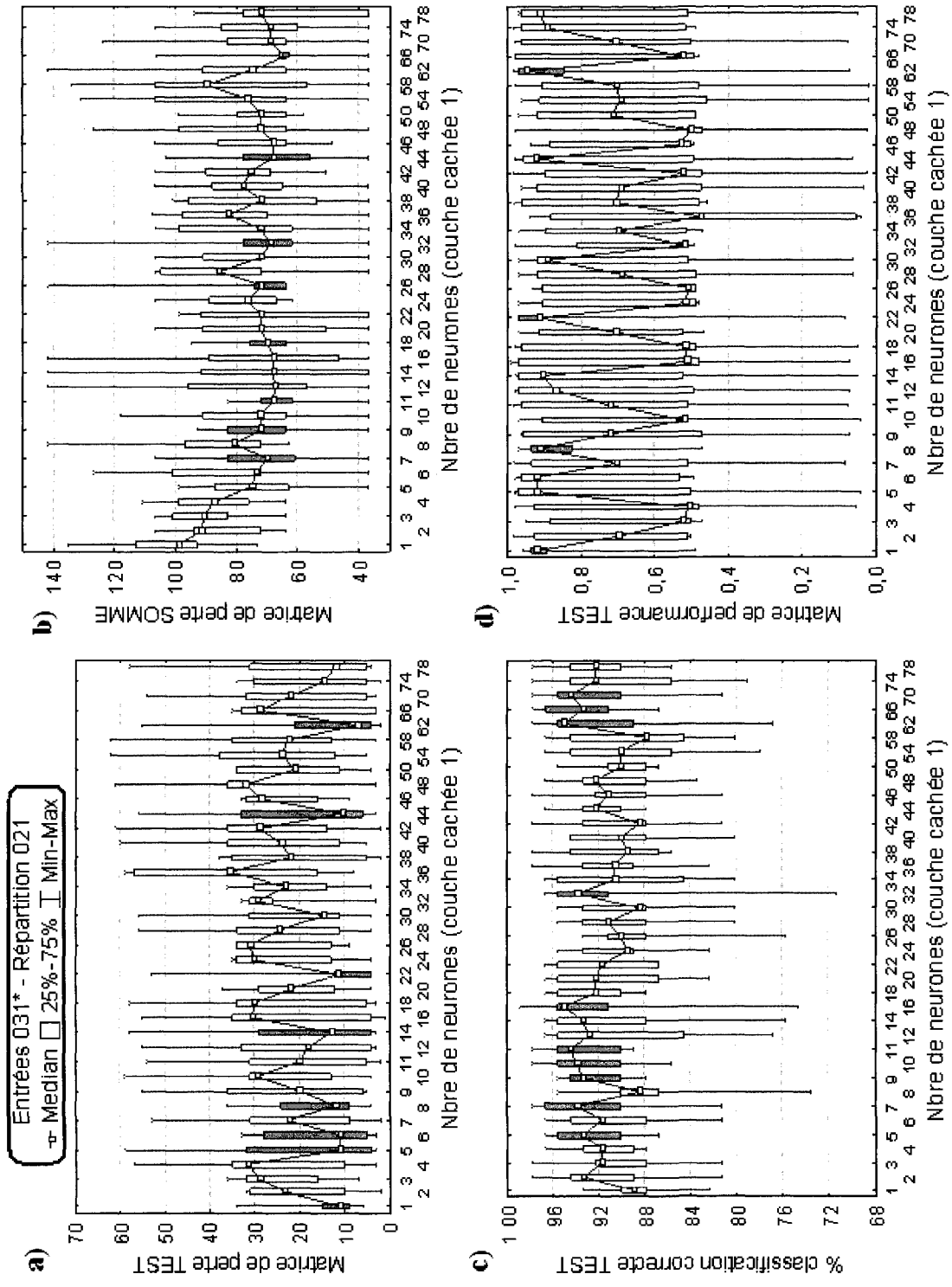

Figure 3-4 : Exemple, quatre critères de performance, Entrées 031*, répartition 021

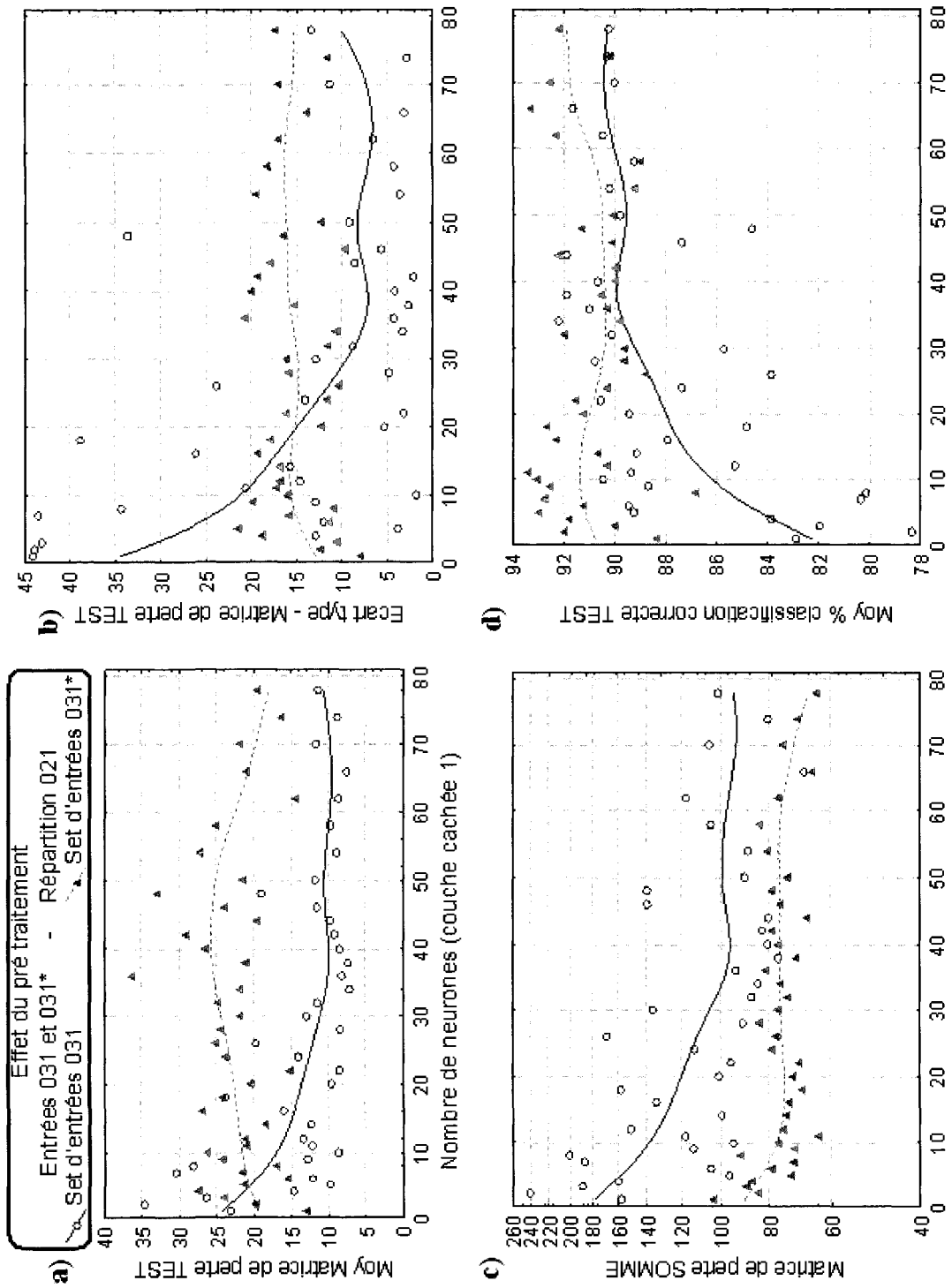

Figure 3-5 : Exemple de l'effet du prétraitement, Entrées 031, répartition 021

Sur la Figure 3-5, sont tracées les valeurs moyennes des dix meilleurs réseaux retenus selon trois critères (matrice de perte TEST et SOMME, et pourcentage de classification correcte TEST), ainsi que l'écart-type pour le critère matrice de perte TEST. Par souci d'allègement, les autres figures pour les autres critères de performance ne seront pas présentés ici, ils fournissent cependant des résultats identiques. Les lignes pleines et en pointillés sont obtenues par l'ajustement des moindres carrés pour les entrées 031 et 031* respectivement.

Nous observons que les entrées 031 (cercles) peuvent atteindre de bons résultats à partir d'un grand nombre de neurones : autour de 40 neurones, la matrice de perte TEST descend autour de 6 ou 7 unités (Figure 3-5 a). Ces résultats dépassent de 9 unités les meilleurs résultats obtenus pour le prétraitement (triangles). Concernant la stabilité (Figure 3-5 b), le prétraitement fournit sur toute la plage de neurones une bonne stabilité comparativement à l'entrée 031 où il faut attendre 24 neurones pour que l'écart-type soit inférieur. Cependant, les entrées 031* se distingue par une meilleure classification à bas nombre de neurones (inférieur à 10 neurones cachés) pour les critères matrice de perte SOMME et pourcentage de classification correcte TEST (Figure 3-5 c et d).

Cet exemple suggère donc l'usage du prétraitement par fonction de répartition pour le seuil 9.3 UTN au printemps. Pour bâtir le modèle final, nous chercherons le meilleur nombre de neurones cachés parmi l'intervalle élargit [6; 10] neurones.

Rajoutons un commentaire sur cet exemple. Bien que présentant des performances légèrement moins bonnes pour la matrice de perte TEST (Figure 3-5 a), le prétraitement présente l'avantage de fournir des performances quasi équivalentes à plus bas nombre de neurones cachés. La recherche d'une solution parcimonieuse est un atout pour prévenir le sur-apprentissage, et pour garantir plus de généralisation du modèle.

Tableaux récapitulatifs des réseaux retenus par saison et par seuil

Les résultats finaux des entrées retenues (prétraitées ou non par fonction de répartition) et du nombre de neurones cachés pour chaque modèle figurent ci-après.

Tableau 3-12 : Réseaux et entrées retenus par seuils - Automne

AUTOMNE					
Seuil (UTN)	**4**	**5,5**	**7,5**	**9,3**	**Régression**
Entrées	TURB_DB-1 RIV_RAIS-1 LSF_VITM-1 LSF_PREC-2 PRECX_DS-1	TURB_DB-1 TURB_HAW-1 RIV_CHAT-1 RIV_CHAT-3 DOR_PREC-2 SAB_VITM-1	DOR_VITX-2 LSF_VITM-1 SAB_VITM-1	TURB_HAW-1 TURB_HAW-2 RIV_CHAT-1 SAB_VITM-1 PRECX_DS-1 PRECX_DS-5	TURB_DB-1 LSF_VITM-1 SAB_VITM-1
Prétraitement ?	Oui	Oui	Non	Non	Non
Meilleure configuration PMC retenue	5:7:1	6:11:1	3:8:1	6:6:1	3:5:1

Tableau 3-13 : Réseaux et entrées retenus par seuils – Eté

ETE					
Seuil (UTN)	**4**	**5,5**	**7,5**	**9,3**	**Régression**
Entrées	En raison du faible nombre d'exemples "hauts", aucun modèle de classification n'a été développé pour l'été.				TURB_DB-1 TURB_DB-3 RIV_CHAT-1 PRECX_DS-1
Prétraitement ?					Non
Meilleure configuration PMC retenue					4:4:1

Tableau 3-14 : Réseaux et entrées retenus par seuils - Printemps

PRINTEMPS					
Seuil (UTN)	**4**	**5,5**	**7,5**	**9,3**	**Régression**
Entrées	TURB_DB-1 RIV_RAIS-6 RIV_CHAT-1 LSF_VITM-1 LSF_PREC-2 IDX_RENV-1	TURB_DB-1 TURB_HAW-3 RIV_CHAT-3 LSF_VITM-1 DOR_PREC-1 LSF_PREC-2 PRECX_DS-3 IDX_FONT-1	TURB_DB-1 OUT_FLV-1 RIV_CHAT-3 DOR_VITM-2 LSF_VITX-2 LSF_VITM-1 DOR_PREC-2 LSF_PREC-2 IDX_FONT-3	TURB_DB-1 TURB_DB-2 OUT_FLV-1 RIV_CHAT-3 RIV_CHAT-5 DOR_VITM-2 LSF_VITM-1 DOR_PREC-2 PRECX_DS-6 IDX_FONT-3	TURB_DB-1 TURB_HAW-3 OUT_FLV-1 RIV_RAIS-6 RIV_CHAT-1 RIV_CHAT-3 LSF_VITM-1 SAB_VITM-1 PRECX_DS-1 PRECX_DS-2 IDX_RENV-1
Prétraitement ?	Non	Non	Oui	Oui	Non
Meilleure configuration PMC retenue	6:2:1	8:22:1	9:9:1	10:8:1	11:8:1

Il est intéressant de remarquer qu'à l'exception du modèle de classification 5,5UTN au printemps demandant 22 neurones cachés, tous les autres modèles nécessitent moins de dix neurones. Ainsi, l'option d'étudier les performances apportées par une deuxième couche cachée ne sera pas considérée étant donné que le nombre de neurones cachés est déjà peu élevé.

3.1.9 Bâtir le modèle final

Mise en commun des modèles de classification

Comme décrit dans la section 2.3.11, il faut calculer une sortie synthèse des modèles de classification mis en cascade, et ce pour chaque saison. Nous codons arbitrairement nos valeurs de sortie tel qu'indiqué dans le Tableau 3-15. Ces valeurs sont choisies de sortes qu'elles soient proches de la médiane des intervalles classifiant. Un « 0 » signifie prédiction « Basse », alors qu'un « 1 » signifie prédiction « Haute ».

Tableau 3-15 : Valeur de sortie des modèles de classification en cascade de TURB_DB

Modèle classifiant au seuil (en UTN)				Valeur de synthèse de la classification (en UTN)
4	5,5	7,5	9,3	
0	*	*	*	Dépend de la saison considérée: Printemps => 2,53 UTN Été => 2,09 UTN Automne => 1,97 UTN
1	0	*	*	5
1	1	0	*	6,5
1	1	1	0	8,5
1	1	1	1	13

* : quelle que soit la combinaison (0 ou 1)

Mise en commun des prédictions par saison

Il résulte donc pour chaque saison deux prédictions : une issue de la synthèse des modèles de classification, et une issue du modèle de régression. Afin de prendre en compte la variabilité des dates séparant une saison à l'autre, un indice de pertinence de chaque saison est créé. Ce nombre varie de 0 à 1, 1 signifiant que la saison considérée est la plus probable. Cet indice de pertinence en fonction de la date est décrit à la Figure 3-6. Les dates frontières entre deux saisons identifiées à la section 3.1.4 figurent en gras.

L'été étant la saison la moins critique en termes de turbidité à l'eau brute à Des Baillets, sa transition sera donc écourtée à +/- 3 jours autour de la date de transition, ceci au profit des autres saisons plus problématiques, à savoir printemps et automne.

L'utilisateur final sera ainsi averti des prédictions fournies par chaque modèle saisonnier par ordre de pertinence. En cas de chevauchement de deux saisons, l'affichage de la deuxième saison potentielle pourrait fournir un avis conservateur

dans la situation où les saisons seraient décalées temporellement. Il peut s'agir par exemple d'une fonte des neiges précoce.

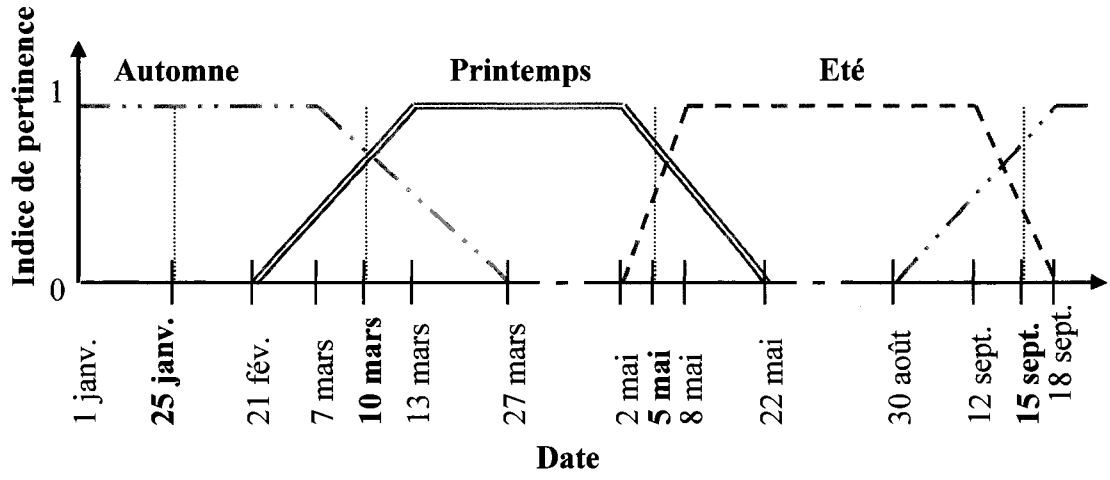

Figure 3-6 : Indice de pertinence des modèles saisonniers DB en fonction de la date

3.2 Résultats

Les exemples de la base de données étant divisées aléatoirement, il n'est pas possible de reconstituer une année entière de données consécutives dans le temps qui n'aient pas servi à l'apprentissage des modèles. Ainsi, nous ne pourrons quantifier la performance du modèle fonctionnant sur une année de test entière. Cependant, les résultats de chaque saison prise indépendamment vont être étudiés.

Le tracé de la turbidité prédite (par classification et par régression) est comparé à celui de la turbidité réellement observée. Pour chaque exemple constituant les ensembles Test des répartitions (x99) de chaque saison afin d'évaluer la performance générale des modèles. Les mêmes figures pour les répartitions (x98) figurent à l'Annexe B.

Un tableau récapitulatif à l'Annexe B résume les critères de performance sur l'ensemble Test et toutes les données de chaque saison. Ces critères étant, pour le modèle de classification, le pourcentage de classification correct de chaque seuil; et

pour le modèle de régression, le coefficient de corrélation, la racine carrée de l'erreur quadratique moyenne (EQM ou RMSE en anglais), et l'erreur absolue moyenne (EAM).

Enfin dans une troisième partie, à titre de comparaison, les pourcentages d'amélioration par rapport au modèle linéaire de référence équivalent (i.e. avec le même groupe d'entrées que celles finalement retenues pour les modèles de type PMC) sont présentés pour chaque seuil de classification, et ce pour les critères de pourcentage de classification correct et de matrice de performance.

Saison : Printemps

Parmi les observations supérieures à 4UTN, seuls les exemples 36 et 38 sont mal classés. Ils correspondent respectivement aux dates et valeurs de turbidité suivantes : 11 mars 2002 (21UTN), et 23 mars 2002 (4,60UTN). Le premier ne présente pas une situation problématique : bien que prédit à une valeur inférieure à celle réellement observée, il ne s'agit pas d'un faux négatif puisque qu'un pic de turbidité est quand même prédit. En effet, les modèles de classification en cascade ont annoncé 6,5UTN, et 8,99UTN pour le modèle de régression. Le deuxième exemple correspond à un épisode de turbidité isolé lors d'une possible fragilisation du couvert de glace, épisode accompagné de pluie à trois jours et cinq jours précédant l'évènement, ainsi que de forts vents deux jours auparavant. Les causes responsables de l'évènement turbide en question ne sont pas nettes à cause de la présence hypothétique du couvert de glace. De plus, la valeur observée de 4,60 UTN est proche de la frontière de 4UTN, ce qui expliquerait la difficulté du réseau à discriminer dans une zone où le chevauchement des classes de turbidité est fort.

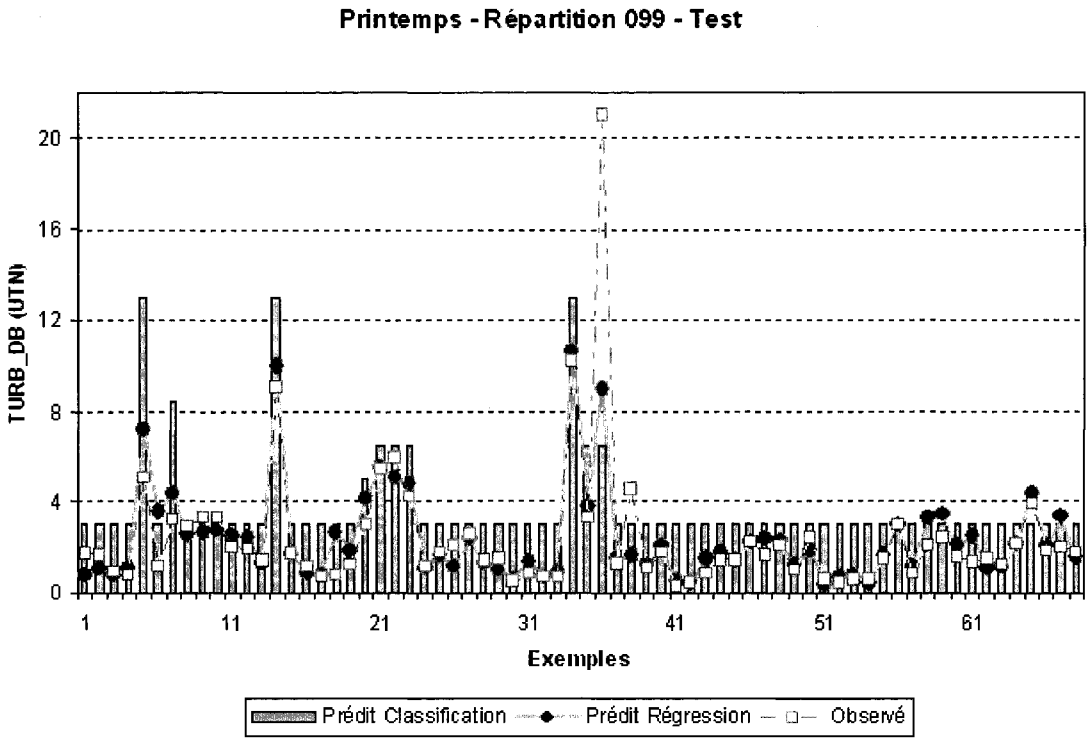

Figure 3-7 : TURB_DB observée et prédite au printemps - r099 – Test

Pour cette figure (Figure 3-7) et ses homologues dans les saisons suivantes, il est important de noter qu'il s'agit bien d'exemples tirés aléatoirement en abscisse et non d'exemples consécutifs dans le temps. Bien que les séries soient reliées entre elles par souci de lisibilité, tous les points sont indépendants les uns des autres.

Tableau 3-16 : Résultats des modèles de classification et régression - printemps

PRINTEMPS	CLASSIFICATION				RÉGRESSION		
	% de classification correcte au seuil				Corrélation	EQM (UTN)	EAM (UTN)
	4	5,5	7,5	9,3			
Test	0,941 (0,911)	0,925 (0,940)	0,956 (0,970)	0,956 (0,925)	0,828 (0,921)	1,66 (1,39)	0,693 (0,702)
Toutes les données	0,927 (0,935)	0,928 (0,946)	0,963 (0,963)	0,971 (0,933)	0,915 (0,919)	1,07 (1,05)	0,548 (0,552)

Au Tableau 3-16 sont indiqués les critères de performance des modèles de classification et de régression pour les ensembles de Test et toutes les données des répartitions 099 et 098 (entre parenthèses).

De très bonnes performances sont atteintes au printemps : concernant les modèles de classification, le pourcentage de classification correcte est toujours supérieur à 91%. Pour la régression, bien que les performances issues de l'échantillonnage 099 soient moins bonnes, un coefficient de corrélation élevé est observé (supérieur à 0,91) pour une erreur quadratique moyenne et une erreur absolue moyenne faibles.

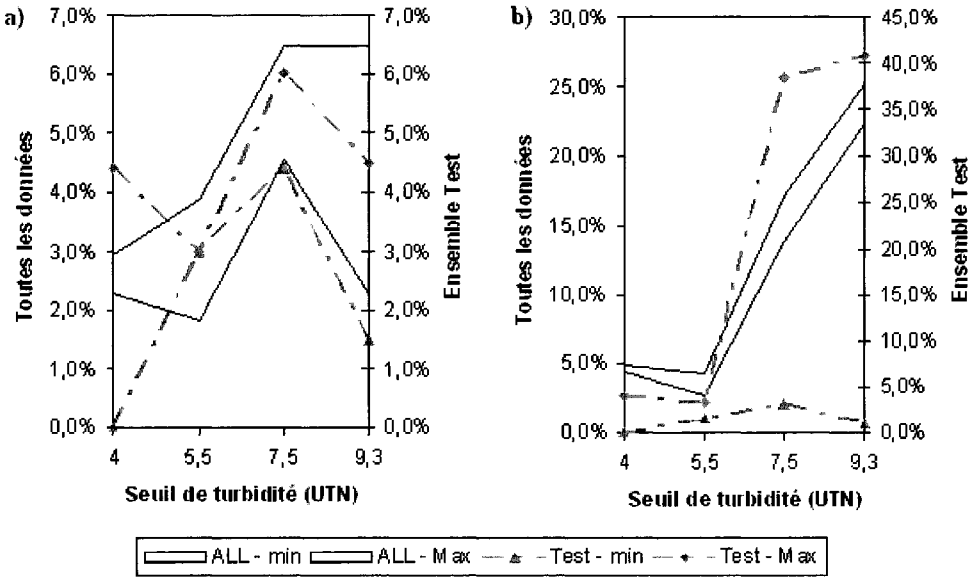

Figure 3-8 : Comparaison PMC - modèle linéaire au printemps. Pourcentage d'amélioration du critère (a) pourcentage de classification correcte, (b) matrice de performance.

Sur la Figure 3-8 est tracée la performance additionnelle (en pourcentage) en fonction du seuil de turbidité classifiant, performance obtenue pour les deux critères suivants : pourcentage de classification correcte (Fig. a) et matrice de performance (Fig. b). Les

courbes « Max » et « min » ont été obtenues à partir des résultats des répartitions 098 et 099.

Il ressort que l'influence sur les données prises dans leur totalité (surface pleine), suit sensiblement la même tendance que les ensembles de Test (traits d'axe).

De plus, le pourcentage de classification est sujet à une légère augmentation des performances par rapport aux modèles linéaires de base (de 0 à 6,5%). Cependant, les fortes augmentations du critère matrice de performance aux seuils 7,5UTN et 9,3UTN (38,4 et 40,8% respectivement pour les ensembles de Test) traduit une nette amélioration des prédictions des évènements hauts. Rappelons-le, le critère matrice de performance a été construit pour accorder deux fois plus d'importance aux évènements de « haute » turbidité. Le faible nombre d'exemples hauts à ces seuils explique la grande variabilité des résultats sur l'ensemble Test (seulement un faux négatif supplémentaire peut grandement dégrader la performance obtenue). Pourtant, l'amélioration de la prédiction à 7,5 et 9,3UTN est confirmée par les gains de performances concentrés autour de 15 et 22% respectivement (quelle que soit la répartition utilisée).

Saison : Été

Aux vues du faible nombre de cas observés supérieurs à 4UTN, ne sera traité ici que le modèle de régression, répertorié dans le Tableau 3-17.

Tableau 3-17 : Résultats des modèles de classification et régression - été

ÉTÉ	CLASSIFICATION				RÉGRESSION		
	% de classification correcte au seuil				Corrélation	EQM (UTN)	EAM (UTN)
	4	5,5	7,5	9,3			
Test	Aucun modèle de classification n'a été développé pour l'été				0,746 (0,792)	0,382 (0,360)	0,254 (0,240)
Toutes les données					0,774 (0,772)	0,370 (0,373)	0,264 (0,267)

Les prédictions sont un peu moins bonnes temporellement pour l'été que pour le printemps (la corrélation passe d'un minimum de 0,91 à 0,74). Elles restent néanmoins précises : du fait des faibles valeurs de turbidité observées l'erreur quadratique moyenne reste très faible. Toutefois, cette faible erreur est aussi liée à la moyenne des valeurs de turbidité à l'été qui est aussi faible (par construction, l'été a regroupé toutes les valeurs inférieures à 5 UTN). L'erreur absolue moyenne (0,25UTN) représente seulement 12% de la valeur moyenne de la turbidité à l'eau brute l'été (2,1 UTN).

Saison : Automne

L'observation de la turbidité prédite et observée pour les 193 exemples de l'ensemble Test r299 (Figure 3-9) révèle une bonne performance de classification générale offerte par les deux modèles (classification ou régression). Seul un évènement turbide n'est pas détecté par l'un ou l'autre des modèles. Il s'agit d'une pointe à 5,30 UTN du 5 octobre 2002 occasionnée probablement par les vents à Dorval le jour même et la veille (DOR_VITM et DOR_VITX 0 et -1). Les paramètres du jour même étant inaccessibles, seul les valeurs de la veille pourraient être incluses. Or, les vitesses du vent maximales et moyennes à Dorval la veille ont déjà été écartées de la liste des variables car moins fréquemment liées à des pointes de turbidité que d'autres variables (comme SAB_VITM-1 par exemple).

De plus, notons que le modèle de classification surestime bien souvent les pointes avec un plus grand nombre de faux positifs que le modèle de régression. Ceci semble logique dans la mesure où le modèle de régression dispose pour sa calibration d'une plus grande proportion d'exemples inférieurs à 5 UTN que de pointes de turbidité.

Tableau 3-18 : Résultats des modèles de classification et régression - automne

AUTOMNE	CLASSIFICATION				RÉGRESSION		
	% de classification correcte au seuil				Corrélation	EQM (UTN)	EAM (UTN)
	4	5,5	7,5	9,3			
Test	0,881 (0,871)	0,819 (0,813)	0,739 (0,803)	0,937 (0,882)	0,797 (0,771)	1,53 (1,52)	0,667 (0,597)
Toutes les données	0,871 (0,872)	0,766 (0,798)	0,760 (0,796)	0,952 (0,873)	0,741 (0,725)	1,60 1,63	0,707 (0,671)

À la lecture du Tableau 3-18, il ressort que les performances obtenues à l'automne sont moins bonnes que celle du printemps, les phénomènes à modéliser y étant sans doute plus complexes. En effet, le modèle de classification affiche des pourcentages de classification correcte plus variables (ils s'étalent de 74% à 95%). La discrimination semble plus aisée seulement pour le seuil 9,3 UTN où les performances observées se démarquent plus nettement des trois autres seuils. Pour la régression, la corrélation affichée est elle aussi plus basse : de l'ordre de 0,78 pour l'ensemble de Test, comparé à 0,87 au printemps. Les valeurs de l'EQM et de l'EAM restent du même ordre de grandeur qu'au printemps.

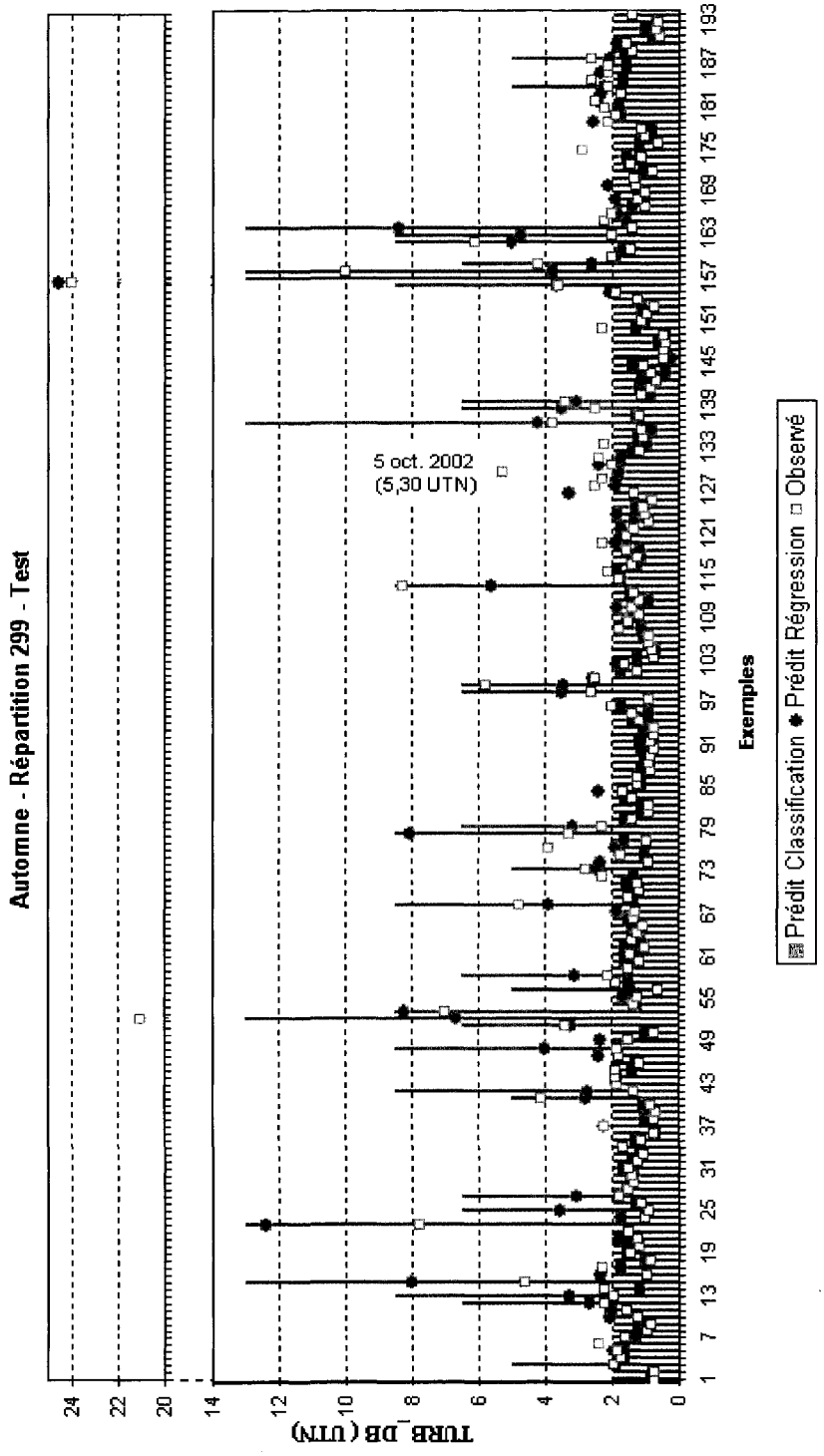

Figure 3-9 : TURB_DB observée et prédite à l'automne- r299 – Test

Sur la Figure 3-10, l'amélioration apportée par le modèle RNA peut atteindre des nombres dont la moyenne est significativement plus élevée qu'au printemps (où le gain de performance du pourcentage de classification correcte s'étalait de 0 à 6,5%). La variabilité des résultats obtenus pour les deux répartitions est aussi plus grande. Il semble qu'à l'automne la discrimination entre les classes hautes et basses soit moins franche qu'au printemps, ainsi les RNA se distinguent par leur capacité à construire des frontières discriminantes plus complexes. Ceci aboutit à une diminution des faux positifs tout en maintenant une performance au moins égale en termes de classification des évènements de « haute » turbidité. Effectivement, les gains de performance obtenus pour le critère pourcentage de classification correcte sont supérieurs aux valeurs des gains pour le critère matrice de performance (critère mettant l'accent sur les évènements hauts).

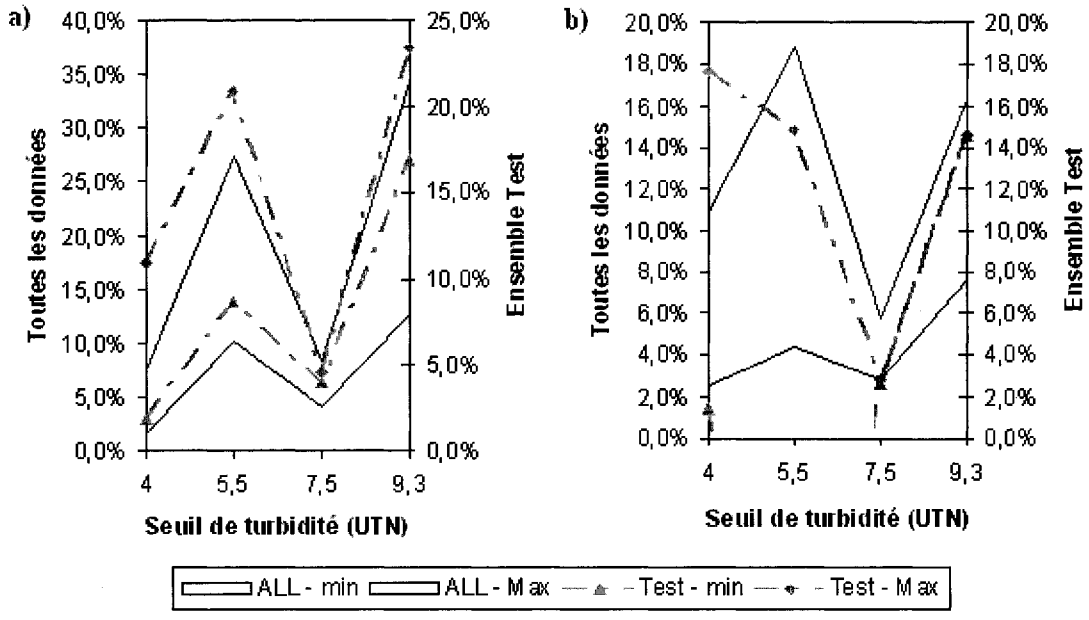

Figure 3-10 : Comparaison PMC - modèle linéaire à l'automne. Pourcentage d'amélioration du critère (a) pourcentage de classification correcte, (b) matrice de performance.

En Figure 3-10 b, le critère matrice de performance suggère une amélioration négative au seuil 5,5UTN pour l'ensemble Test de la répartition 298. Ceci s'explique par le fait que le modèle choisi contient deux faux négatifs de plus que son équivalent linéaire, au bénéfice d'une diminution des faux positifs. Ce choix du modélisateur est justifié pour éviter le sur-apprentissage sur des exemples turbides spécifiques, surtout lorsque ces exemples sont situés dans une zone où les classes de turbidité se chevauchent fortement. Au prix d'un ou deux faux négatifs supplémentaires de l'ensemble Test, nous avons privilégié une très nette diminution des faux positifs (jusqu'à 35% de classification correcte supplémentaire en Figure 3-10 a).

Les seuils 7,5 et 9,3UTN semblent plus aisés à discriminer: en effet, les gains de performance sont moins sensibles à la répartition utilisée. Ceci ce vérifie graphiquement, ils correspondent à des zones où il y a peu de chevauchement des classes de turbidité.

De plus fortes valeurs de l'EQM sont observées pour le printemps et l'automne que pour l'été (jusqu'à quatre fois plus grandes). Ceci vient du fait que le calcul de l'EQM se base sur l'espérance du carré de l'erreur. L'espérance a été calculée avec la moyenne arithmétique. Or, cette moyenne est relativement sensible aux valeurs extrêmes surtout présents au printemps et à l'automne.

3.3 Discussion

3.3.1 Données supplémentaires pour améliorer les prédictions

Afin d'augmenter les performances accessibles, il serait intéressant d'inclure des variables de qualité de l'eau brute directement en amont de la prise d'eau. Ceci permettrait d'avoir une meilleure idée de la qualité de l'eau à l'entrée du lac Saint-Louis. Ces variables de qualité de l'eau seraient la turbidité à l'eau brute pour les stations de filtration de Beauharnois et de Vaudreuil. La première reflèterait la qualité

de l'eau du fleuve, alors que la seconde donnerait une information sur la qualité de l'eau en provenance du lac des Deux Montagnes, eau passant par la gire de l'île Perrot. La qualité de l'eau à Beauharnois n'est disponible seulement depuis le 1er janvier 2002, c'est pourquoi elle fut écartée dans ce modèle. Pour Vaudreuil, les données existent mais n'ont pas pu être récupérées pour ce projet.

3.3.2 Commentaires sur l'usage du prétraitement par fonction de répartition

Le but de ce prétraitement est d'utiliser l'information préalablement connue sur la distribution d'une variable d'entrée donnée. Une amélioration des prévisions obtenues n'est observée que pour deux des quatre modèles de classification.

Ces performances ne sont pas observées dans tous les modèles. À cause de la faible proportion d'évènements extrêmes dans les bases de données, les transformations écrasent ces évènements dans des intervalles restreints. Par exemple, sur la Figure A-5, les valeurs de TURB_DB-1 supérieures à 4 UTN sont toutes comprimées dans l'intervalle [0,84 ; 1]. Cet écrasement des données agirait de manière conservatrice dans des cas où la classification est peu aisée : les faibles valeurs de turbidité à l'automne, et les hautes turbidités au printemps. Ils amélioreraient la prédiction des évènements hauts au prix d'un plus grand nombre de faux positifs.

3.3.3 Commentaires sur la méthode de sélection des entrées

Il est important de noter que les méthodes de sélection des entrées ne donnent aucun résultat absolu. En effet, selon la répartition des exemples et les méthodes utilisées (GAPNN, « *forward stepwise* », « *backward stepwise* », etc.), les résultats et les entrées candidates choisies ne sont pas les mêmes. Parmi les entrées non élaguées par les méthodes de sélection, le choix final des entrées à retenir par modèle est laissé à la disposition du modélisateur.

Ainsi, tous les efforts effectués au préalable permettent une meilleure compréhension des phénomènes explicatifs récurrents et oriente la sélection finale d'entrées. Ces travaux sont l'observation graphique du phénomène à modéliser, le recensement des évènements turbides, le tracé des diagrammes de points catégorisés (afin d'observer le pouvoir séparateur de deux variables descriptives), etc.

Après avoir eu recours aux méthodes énoncées dans la Section 2.3.5 pour éliminer le plus grand nombre de variables non pertinentes, il reste une série d'entrées dont l'information peut être redondante dans certains cas. Le choix final des entrées s'est fait à l'aide de ces connaissances préalables et d'essais et erreurs sur l'introduction de variables explicatives supplémentaires.

Cette redondance d'information et la corrélation existant entre plusieurs variables potentiellement candidates est responsable en partie de la variabilité des résultats obtenus quant à la sélection des entrées pertinentes. Par exemple, les méthodes employées peuvent retenir indifféremment le vent le même jour à deux places différentes ou bien des variables qui sont la conséquence d'autres variables (au printemps la fonte des neiges implique la hausse du débit de tributaires secondaires).

Par conséquent, dans le but d'obtenir des sélections plus franches des variables d'entrées pertinentes, il conviendrait de réduire l'espace des entrées et de s'assurer que ces dernières soient non corrélées entre elles. L'analyse en composantes principales (ACP) est la technique la plus couramment employée pour projeter orthogonalement les variables dans une nouvelle base maximisant la variance des variables. Cependant, cette technique induit une perte d'information lors de la projection orthogonale (Dreyfus et al., 2004). Une autre technique à explorer serait l'utilisation de cartes auto-organisatrices (Bowden et al., 2005).

Par souci de simplification et pour ne pas perdre d'information contenue dans les données, ces méthodes n'ont pas été utilisées ici. Si elles étaient utilisées, une

sélection des entrées plus parcimonieuse que celle obtenue ici pourrait être accessible. Ceci permettrait d'augmenter la performance de généralisation du modèle.

Chapitre 4 PRÉDICTION DE LA TURBIDITÉ À LA PRISE D'EAU BRUTE À LA STATION ATWATER

4.1 Étapes de la modélisation

4.1.1 Définition des objectifs et mise en contexte

L'objectif de ce chapitre est de prédire la turbidité à l'eau brute, une journée à l'avance, pour l'usine Atwater (variable TURB_ATW). Cette dernière puise son eau de la même source d'eau brute que la station Charles J. Des Baillets, à savoir le fleuve Saint-Laurent, et est située en aval du canal Atwater (Figure 4-1).

L'usine Des Baillets est directement alimentée par l'eau provenant de la prise d'eau brute. L'usine Atwater puise son eau à l'extrémité du canal. Huit kilomètres de canal à ciel ouvert séparent ainsi les deux usines. Le canal a une section de 245 m^2 (49m x 5m). L'usine filtre en moyenne annuelle environ 496 000 m^3/d (données de 2006), ce qui fait un temps de séjour moyen dans le canal de quatre jours. En période de forte production, ce temps peut être réduit à trois jours et moins. En termes d'occupation des sols, les abords du canal sont principalement constitués d'agglomérations, de quelques terrains de sports et de parcs (parc Angrignon). Au bout du canal, sur la rive nord, l'autoroute A15 longe la berge juste au devant de l'usine Atwater (voir Figure 4-1).

Par conséquent, l'usine Atwater est sous influence directe de la qualité de l'eau à la prise d'eau qui alimente l'usine des Baillets. Le canal pourrait servir de décanteur et ainsi diminuer la turbidité à la prise d'eau. Cependant, le fait que ce dernier soit à ciel ouvert peut aussi le rendre vulnérable aux précipitations et ruissellements urbains. Les causes potentielles pouvant affecter TURB_ATW sont donc la qualité de l'eau en amont et les jours précédents; et dans une moindre mesure, l'apport externe par ruissellement lors de la fonte des neiges ou de fortes précipitations locales.

Dans ce cas particulier, il se pourrait qu'un modèle linéaire simple soit amplement suffisant pour prédire la turbidité à Atwater en fonction de l'eau brute reçue à Des Baillets les jours précédents. Un gain de complexité pourra être atteint en examinant la réponse du canal au fil des saisons, et un découpage en saisons serait à envisager ici aussi.

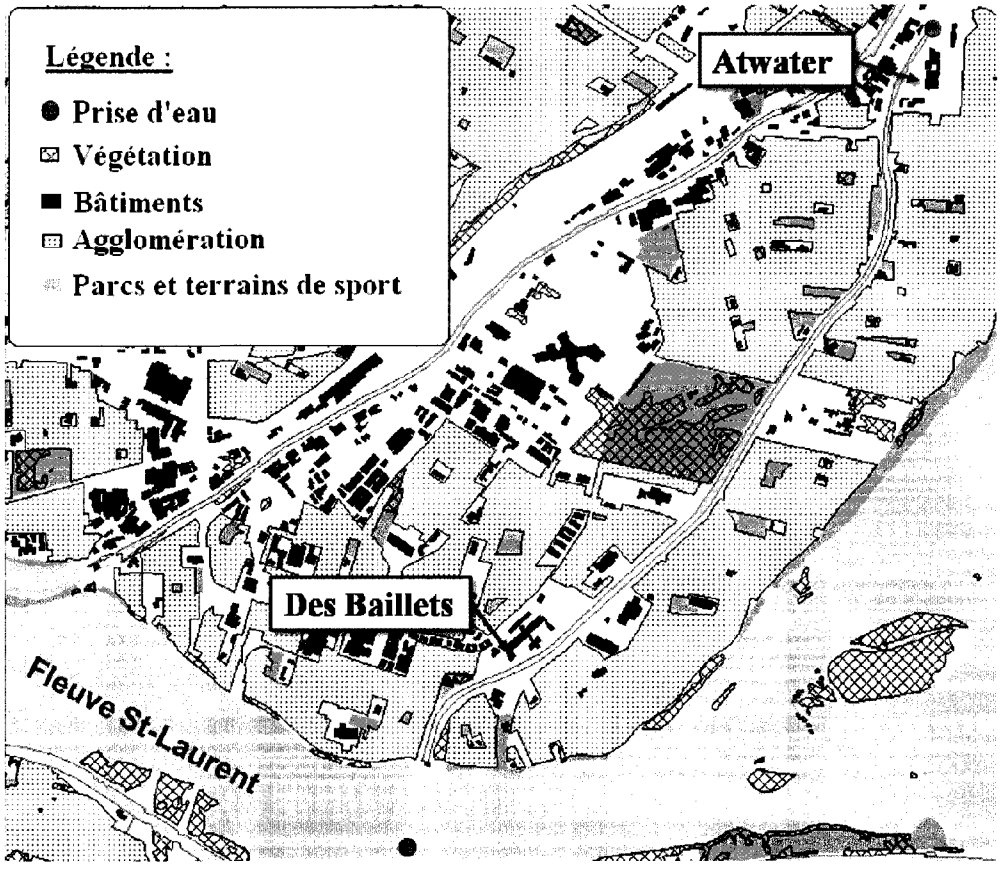

Figure 4-1 : Plan d'occupation des sols autour du canal Atwater (source : Ressources Naturelles Canada)

4.1.2 Récupération des données et inspection

Les données ont déjà été récupérées et triées pour le chapitre précédent. La base de données s'étale du 1er janvier 1996 au 31 mai 2006. Soit environ 3760 données

moyennes journalières exploitables. Un tri fut effectué précédemment mais l'observation de la turbidité à l'eau brute à Atwater (TURB_ATW) en fonction de la date julienne révèle deux valeurs aberrantes qui furent enlevées. Le graphe résultant est donné à la Figure 4-2.

En effet, deux valeurs semblèrent aberrantes : une valeur de 7 UTN le 11 février 1996 et de 8 UTN le 7 septembre 1998. Les valeurs de turbidité observées autour de ces points sont toutes inférieures à 1 UTN, ces brusques pointes créent des discontinuités dans les mesures. Ils ne sont pas précédés de pointes à DB les jours d'avant, ni de précipitations abondantes (>5mm à Dorval ou Sainte-Anne de Bellevue). Aucune des causes potentielles n'expliquent ces pointes, il pourrait s'agir d'un facteur extérieur non pris en compte ici. Ces pointes ont été remplacées par la valeur moyenne du jour précédent et du jour suivant la mesure.

Sur la Figure 4-2, un comportement saisonnier se distingue de nouveau. Les trois saisons (automne, printemps, et été) sont marquées par une bonne qualité moyenne de l'eau, l'automne et l'été sont peu variables alors que le printemps montre de forts pointes de durées prolongées. Une série de pointes isolées viennent ponctuer les trois saisons. Les pointes du printemps sont les plus forts : ils peuvent monter jusqu'à 15,6UTN. Ensuite viennent les pointes de l'automne de bien moindre intensité (jusqu'à 6,3UTN).

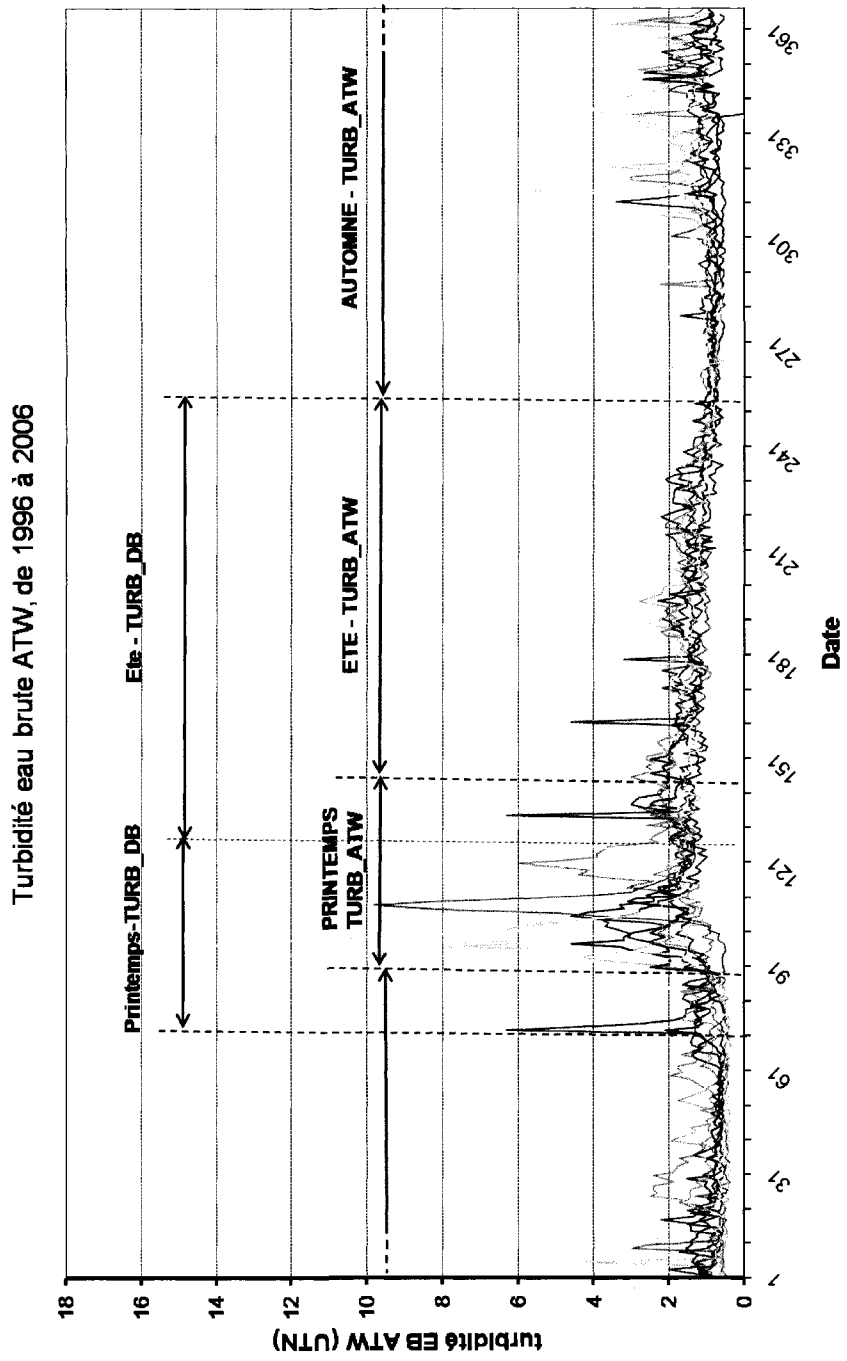

Figure 4-2 : TURB_ATW en fonction de la date julienne de 1996 à 2006

Vers un nouveau découpage temporel ?

Un léger décalage d'une vingtaine de jours semble exister entre les saisons identifiées pour l'usine Des Baillets et celles issues de l'inspection visuelle de la turbidité en fonction du temps (Figure 4-2). Pour chaque saison, les limites temporelles seraient donc à redéfinir pour classer les valeurs de turbidité à l'usine Atwater. On note les observations générales suivantes :

- Le printemps durerait un peu plus longtemps qu'à Des Baillets, jusqu'au j142, soit le 22 mai. Pour la date de début, cela dépend des causes responsables du pic de 6,3UTN le 13 mars 2002. S'il s'agit de la fonte des neiges, ce sera un phénomène dit printanier, s'il s'agit de mauvaise qualité de l'eau en amont, ce sera un phénomène automnal. C'est cette deuxième explication qui prévaut : l'indice de fonte des neiges n'est actif qu'à partir du 30 mars 2002, et un pic de turbidité de 21 UTN est enregistré deux jours avant à Des Baillets. Donc, le printemps à Atwater aura pour dates : du 25 mars au 22 mai.

- Par souci de simplification, l'été et l'automne pourraient garder les mêmes dates, soit respectivement du 23 mai au 15 septembre, et pour l'automne du 16 septembre au 24 mars.

Si l'analyse des entrées suggère un modèle par saison, c'est ce nouveau découpage temporel qui sera adopté pour la prévision d'Atwater.

4.1.3 Analyse statistique de TURB_ATW

Les statistiques descriptives sont données au Tableau 4-1 pour toutes les données, et pour chaque saison.

Tableau 4-1 : Statistiques descriptives TURB_ATW

	TURB_ATW (en UTN)									
	N	Moyenne	Médiane	Minimum	Maximum	P25	P75	P90	P95	Ecart type
Toutes les données	3772	1,27	1,1	0,33	15,6	0,8	1,5	1,97	2,4	0,89
Printemps	628	2,1	1,7	0,50	15,6	1,35	2,3	3,4	4,6	1,6
Eté	1163	1,3	1,31	0,38	4,6	1,05	1,6	1,87	2	0,42
Automne	1981	0,97	0,85	0,33	6,3	0,66	1,12	1,48	1,8	0,50

Sur toute l'année

Sur toute l'année (3772 exemples au total), la moyenne est relativement faible (1,27 UTN), le maximum est atteint le 5 avril 1998, avec 15,6UTN. Les données sont relativement concentrées autour de faibles valeurs de turbidité : la médiane (P50) à 1,10 UTN, le 90ème centile à 2,00 UTN environ, le 95ème centile à 2,40 UTN ; et le 99ème centile à 4,60 UTN environ. Globalement, une bonne qualité générale de l'eau est observée (95ème centile seulement à 2,40 UTN).

Par saisons

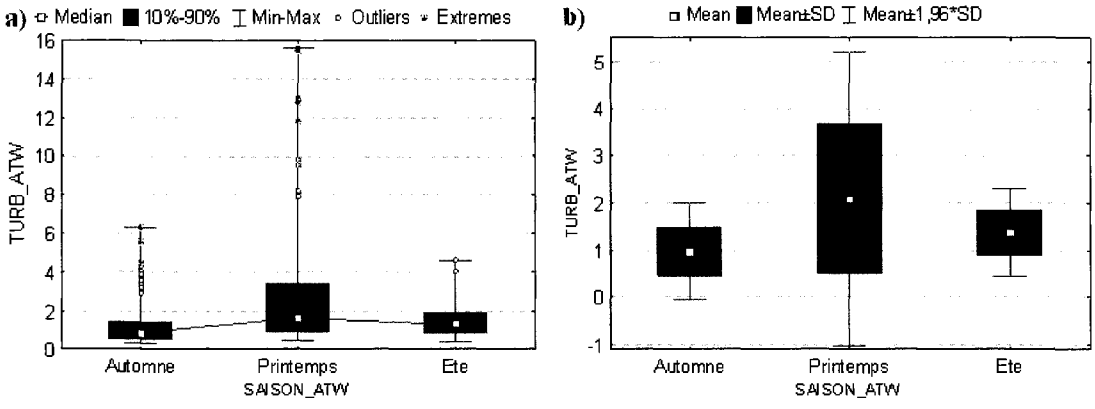

Figure 4-3 : Diagrammes des turbidités de type boîtes à moustaches par saison de TURB_ATW. (a) Médiane et centiles. (b) Moyenne et écart-type

Le printemps montre une moyenne et une variance plus élevée (Tableau 4-1 et Figure 4-3 b). L'automne et l'été présentent peu de variance et des valeurs moyennes quasi identiques, bien que l'été montre une qualité générale de l'eau moindre. En termes de centiles (Figure 4-3 a), 90% des données de l'automne et de l'été sont en deçà de 2 UTN, contre 68% au printemps.

L'été présente moins de pointes de turbidité avec seulement deux pointes observées. Puis vient l'automne, et finalement le printemps avec le plus grand nombre de pointes observés.

Concernant l'intensité de ces pointes, le printemps arrive en tête avec des pointes variant de 7,9 UTN à 15,6 UTN, puis vient l'automne (jusqu'à 6,3 UTN) et finalement l'été. L'amplitude saisonnière observée ici est contraire à celle de l'eau brute à Des Baillets où l'automne présentait des pointes de plus fortes intensités que le printemps.

Il semble qu'il y ait encore deux saisons majeures qui régissent ces pointes : l'automne et le printemps. Comme vu précédemment, la fonte des neiges au printemps entraîne un rapport de mélange accru entre la rivière des Outaouais et le fleuve Saint-Laurent, la qualité générale de l'eau se dégrade donc. Sa couleur se dégrade aussi, elle est plus chargée en particules colloïdales. À l'automne, les pointes observées pour Des Baillets sont principalement dus à l'influence de tempêtes de vents remettant en suspension les sédiments. Ces derniers peuvent décanter dans le canal Atwater, et ne doivent pas être du même type de particules qu'au printemps, sans doute des particules de plus grosse taille. Ceci expliquerait l'inversion de tendance quant à l'intensité des pointes saisonnières.

Si les modèles de régression capturent mal les évènements extrêmes, ceux-ci étant trop peu nombreux ou trop différents du reste des données, un modèle de classification devrait être créé. Ce modèle serait spécialisé dans la détection de pointes de turbidité. Le seuil de coupure adopté doit être un bon compromis : s'il est trop bas la séparation

avec le reste des données est difficile, s'il est trop haut, il n'aura pas assez d'exemples de « haute » turbidité. Le seuil de coupure 4 UTN pourrait convenir pour prédire les pointes printanières. Il est en deçà de la valeur de 5 UTN, où la turbidité de l'eau brute pourrait présenter un risque pour la filtration (en supposant que la filtration sur sable enlève 80% de la turbidité à l'eau brute).

Un modèle de classification en automne n'est pas développé car trop peu d'exemples « hauts » sont disponibles (sept exemples supérieurs ou égaux à 4 UTN). De plus, les quelques exemples 'hauts' ne représentent pas de risque pour le traitement (turbidité maximale de 5,6 et 6,3 UTN), et ils sont prévisibles par des causes évidentes. En effet, la première hausse majeure (5,6 UTN) arrive deux jours après un pic de 31,3 UTN à DB (le 8 novembre 2005); et la deuxième, le 13 mars 2002, est précédée d'un pic de 21 UTN à Des Baillets (le 11 mars).

4.1.4 Partitionnement des exemples

Deux répartitions des exemples sont créées sur le principe de l'échantillonnage aléatoire à proportion fixe (voir chapitre précédent). Pour chaque saison, et pour chaque exemple des classes inférieures à 4 UTN et supérieures ou égales à 4 UTN, une proportion fixe de 70% – 15% – 15 % est répartie aléatoirement entre les ensembles Train – Select et Test respectivement. Cette répartition aléatoire est répétée jusqu'à ce que le test de Kruskall-Wallis et l'observation du diagramme boîtes à moustaches par saison renforce l'idée que les trois ensembles sont extraits de la même population statistique. Ces variables de répartition sont nommées « RepartitionATW1 » et « RepartitionATW2 ».

4.1.5 Choix des variables d'entrées

L'observation graphique des causes actives au voisinage des évènements turbides a déjà été menée (Chapitre 3).

Analyse de la corrélation

Une analyse statistique révèle l'importance du facteur de qualité de l'eau brute à Des Baillets les jours précédents (Tableau 4-2). Le maximum de corrélation par type de données est indiqué en gras.

Tableau 4-2 : Corrélations croisées de TURB_ATW avec les données de qualité de l'eau en amont

	TURB_ATW			
	Toute l'année	Automne	Eté	Printemps
TURB_DB	0,50	0,32	**0,28**	0,77
TURB_DB-1	0,57	0,48	0,24	0,83
TURB_DB-2	**0,65**	**0,65**	0,23	**0,89**
TURB_DB-3	0,56	0,44	0,23	0,83
TURB_HAW-1	0,53	0,44	0,25	0,47
TURB_HAW-2	0,58	**0,46**	0,27	0,55
TURB_HAW-3	**0,62**	0,43	**0,31**	**0,63**
TURB_ATW-1	**0,88**	**0,71**	**0,79**	**0,89**
TURB_ATW-2	0,79	0,59	0,74	0,78
TURB_ATW-3	0,71	0,53	0,73	0,65

L'analyse de la corrélation de TURB_ATW en fonction des paramètres de qualité amont et les jours précédents montre que la turbidité de l'eau brute d'Atwater est fortement corrélée avec la turbidité à Des Baillets deux jours avant (r = 0,89 au printemps et 0,65 sur toute l'année). Le décalage de deux jours représente le maximum de corrélation pour les trois saisons et sur toute l'année. Pour les données d'Atwater, le décalage j-1 est prépondérant (r varie de 0,71 à 0,89 au printemps). Les plus hautes corrélations sont obtenues pour le printemps.

Un modèle annuel ou saisonnier simple comprenant TURB_DB-2 et TURB_ATW-1 comme descripteurs suffirait pour obtenir de bons résultats. Toutefois, au printemps, TURB_DB-2 et TURB_ATW-1 sont fortement auto-corrélées (r = 0,83). Les modèles linéaires pourraient être suffisants pour obtenir de bonnes prédictions.

Recours aux modèles saisonniers ?

Le recours aux modèles saisonniers sera adopté si l'inclusion de la variable IDX_SAISON_ATW s'avère être un bon prédicteur pour le modèle annuel.

Pour le modèle annuel, les entrées pertinentes sont sélectionnées à l'aide de réseaux de type GRNN (voir section 3.1.4). Pour chaque répartition, les variables candidates sont : TURB_DB-1-2-3, TURB_ATW-1-2-3, et IDX_SAISON_ATW.

Après l'analyse exhaustive de toutes les combinaisons des variables d'entrées, il ressort que la variable IDX_SAISON _ATW est sélectionnée à chaque fois. L'élaboration de modèles saisonniers serait donc une bonne idée. De plus, TURB_ATW-1 et TURB_DB-2 reviennent aussi régulièrement.

Par conséquent, trois modèles de régression (un par saison) seront développés.

Détermination des entrées candidates par GRNN

La liste des entrées pertinentes sera déterminée par GRNN. Les variables candidates fixes sont, pour chaque saison, TURB_ATW-1 et TURB_DB-2. Elles sont accompagnées des variables représentant les causes majeures identifiées pour la prédiction saisonnière de l'eau brute à Des Baillets. La fenêtre temporelle recherchée est augmentée de deux journées en raison de la forte corrélation entre la turbidité de l'eau à Atwater le jour même et celle de Des Baillets deux jours auparavant (Tableau 4-2). Les principales variables candidates sont rappelées ci-après (Tableau 4-3).

Tableau 4-3 : Variables candidates secondaires pour la prédiction de TURB_ATW

Cause explicative des forts pics de turbidité	**Saison (P=Printemps; A=Automne)**	**Variables explicatives principales pour prédire TURB_DB**	**Décalage temporel exploré pour prédire TURB_ATW (en jours)**
Contribution des Outaouais	P	OUT_FLV-1	de -1 à -3
Renversement	P	IDX_RENV-1	de -1 à -4
Fonte des neiges	P	IDX_FONT-1 à -3	de -1 à -5
Tributaires secondaires	P	RIV_CHAT-1 à -5	de -1 à -7
Vent	P et A	LSF_VITM-1	de -1 à -3
		SAB_VITM-1	de -1 à -3
Précipitation	A	DOR_PREC-2	de -1 à -4

Note : les précipitations de 1 à 4 jours ont été incluses pour représenter l'hypothèse du ruissellement urbain à proximité du canal traversant la ville.

Les modèles candidats sont de type GRNN, il y a trois méthodes de sélection des combinaisons des entrées : approche constructive (« *forward* »), par élimination (« *backward* »), et par algorithmes génétiques. Les résultats de ces analyses figurent ci-après (Tableau 4-4).

Tableau 4-4 : Résultats de la sélection des entrées par réseau GRNN

	Régression	
	Printemps	Automne
TURB_DB-1		
TURB_DB-2	FBG	FBG
TURB_DB-3		
TURB_ATW-1	FBG	FBG
TURB_ATW-2		
TURB_ATW-3		
OUT-FLV-1	FBG	
OUT-FLV-2		
OUT-FLV-3	FBG	
RIV_CHAT-1		
RIV_CHAT-2		
RIV_CHAT-3		
RIV_CHAT-4	B	
RIV_CHAT-5	FBG	
RIV_CHAT-6	FBG	
DOR_PREC-1	FBG	
DOR_PREC-2	G	
DOR_PREC-3	G	
DOR_PREC-4		F
DOR_VITX-1		
DOR_VITX-2		FG
DOR_VITX-3		FBG
LSF_VITM-1		
LSF_VITM-2		FBG
LSF_VITM-3		FBG
SAB_VITM-1		
SAB_VITM-2		FBG
SAB_VITM-3	B	FBG
IDX_RENV-1	G	
IDX_RENV-2	FG	
IDX_RENV-3	F	
IDX_RENV-4	B	
IDX_FONT-1	FBG	
IDX_FONT-2	FBG	
IDX_FONT-3	FBG	
IDX_FONT-4	FBG	

Légende :
F : "forward stepwise"
B : "backward stepwise"
G : algorithmes génétiques

À l'automne et au printemps, TURB_ATW-1 et TURB_DB-2 ressortent systématiquement (pour chaque méthode et chaque répartition).

De plus, à l'automne, la variable LSF_VITM-3 se distingue rapidement. Elle est accompagnée de variables redondantes (SAB_VITM-3 et DOR_VITX-3). Les

précipitations (DOR_PREC-4) ont possiblement un impact sur la prédiction car elles ne sont apparues que pour la répartition 2.

Au printemps, les indicateurs de fonte des neiges reviennent principalement, soit par ordre de priorité, IDX_FONT-4 -3, OUT_FLV-4 ou 1, RIV_CHAT-5 et -6. Dans une moindre mesure, les vents et précipitations pourraient avoir une influence LSF_VITM-2 ou -3, et DOR_PREC-1 et -2.

En conclusion, la plupart des entrées révélées ici ont déjà été utilisées pour la prédiction de l'eau brute à Des Baillets, variables accompagnées du décalage temporel de deux jours. Il y a donc risque de redondance de l'information, ces entrées seront ajoutées si et seulement si elles expliquent des pointes non prédits par les modèles incluant juste TURB_DB-2 et TURB_ATW-1.

4.1.6 Méthode suggérée : approche constructive

Une série de modèles de régression sont élaborés avec Statistica, et leurs performances sont comparées. Les réseaux de neurones sont développés avec l'analyse IPS de Statistica. Les paramètres et l'architecture utilisés sont les mêmes que pour le chapitre précédent.

Les critères de performance considérés sont par ordre d'importance, pour les deux répartitions et pour l'ensemble de test et toutes les données : le coefficient de corrélation, l'erreur quadratique moyenne (EQM), et l'erreur absolue moyenne (EAM). L'EQM met plus d'emphase sur les pointes de turbidité mal prédits, elle ne sera calculée que pour le printemps et toutes les données.

La démarche se déroule en quatre étapes (Figure 4-4).

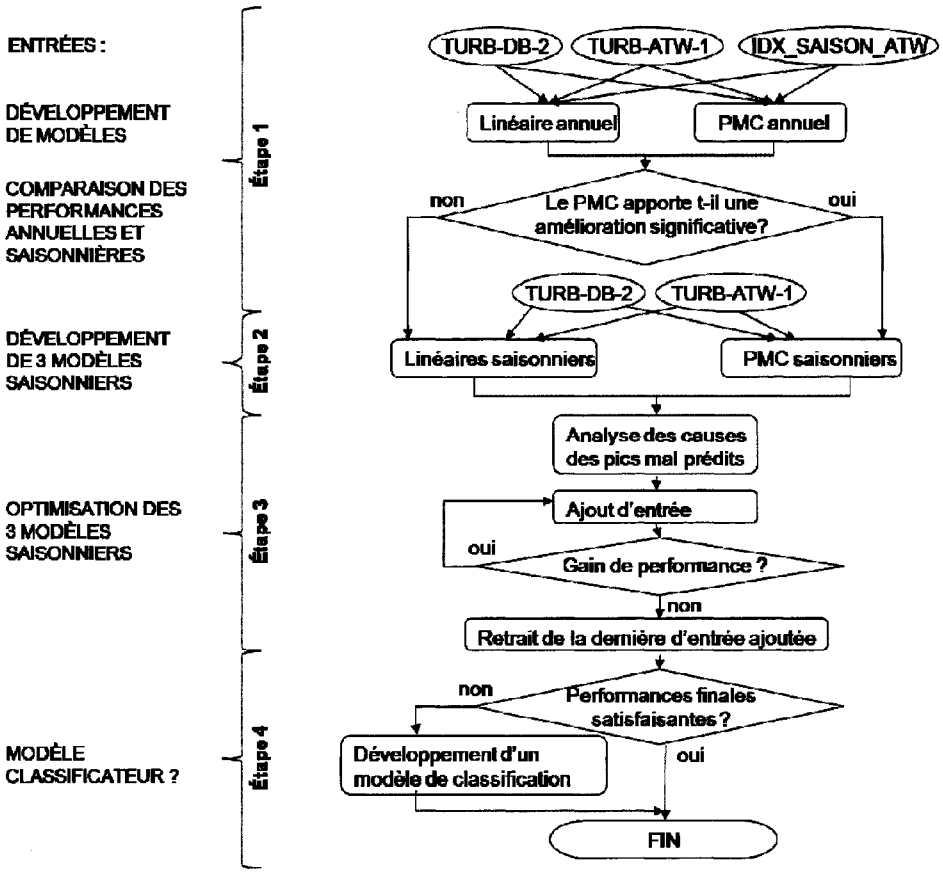

Figure 4-4 : Schéma de la méthode d'élaboration du modèle prévisionnel TURB_ATW

- Étape 1 : modèle de régression pour toute l'année. Pour le modèle linéaire et le PMC, les entrées sont : TURB_DB-2, TURB_ATW-1, et IDX_SAISON_ATW. La performance pour toutes les données et par saison est calculée. Si le modèle neuronal de type PMC n'apporte pas de différence significative dans la performance obtenue, le modèle linéaire sera retenu.

- Étape 2 : découpage en saisons. Trois modèles de régression sont élaborés. Dépendamment des résultats de l'étape 1, les modèles explorés seront linéaires ou neuronaux. Les entrées de base pour chaque saison seront : TURB_DB-2 et TURB_ATW-1. Les résultats des modèles saisonniers sont comparés au modèle annuel.

- Étape 3 (facultative) : analyse des causes potentielles des pointes mal prédits dans les modèles précédents ; au besoin, inclusion de certaines variables explicatives supplémentaires. Est-ce que les variables incluses ont amélioré la performance du modèle ? Si oui, elles sont conservées, sinon retour à l'étape 2.

- Étape 4 : selon les résultats finaux, si les évènements de pointe ne sont pas correctement prédits par les modèles retenus aux étapes 2 et 3, alors l'adjonction d'un modèle de classification doit être envisagée.

4.1.7 Mise en commun des modèles saisonniers

De la même manière que pour l'usine Des Baillets, une probabilité de pertinence des modèles saisonniers est créée. En fonction de la date, cet indicateur avertit l'opérateur du modèle saisonnier auquel il peut accorder le plus de confiance. Si plusieurs modèles se chevauchent, les prédictions de chacun sont affichées par ordre de pertinence croissante. La délimitation des dates est basée sur le découpage saisonnier pour Atwater à la Section 4.1.2, ainsi que sur les hypothèses de la Section 3.1.9. Il en résulte le graphe de l'indice de pertinence des modèles saisonniers pour Atwater en fonction de la date à la Figure 4-5.

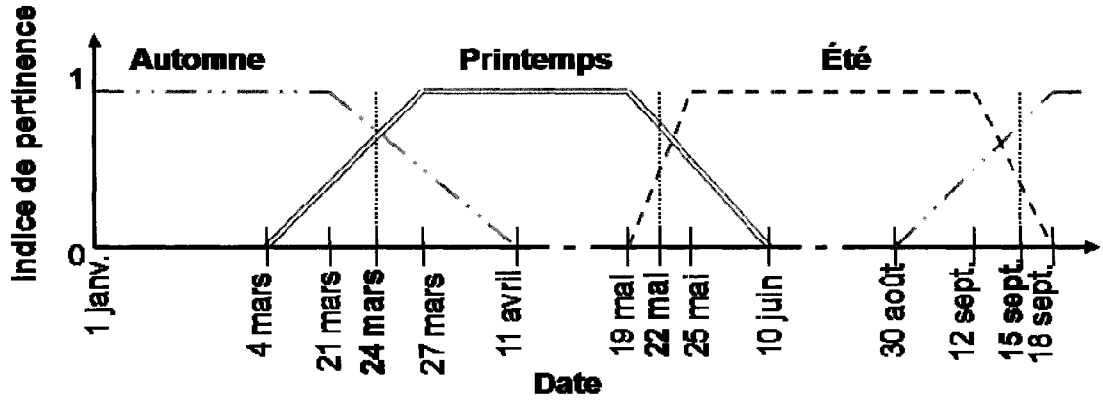

Figure 4-5 : Indice de pertinence des modèles saisonniers ATW en fonction de la date

4.2 Résultats

4.2.1 Étape 1 : modèles annuels

La géométrie du RNA étant déterminée par analyse IPS, la plage de neurones cachés donnant de bons résultats est 14 – 23 neurones. Un faible nombre de neurones permet d'atteindre une plus faible erreur absolue moyenne au prix d'une moins bonne corrélation. Ainsi, un compromis est obtenu avec 17 neurones dans la couche cachée. Le réseau retenu pour les deux répartitions est de type PMC 5 :17 :1. Il y a cinq entrées, au lieu de trois, car la variable catégorielle IDX_SAISON_ATW se retrouve automatiquement codée en plusieurs variables binaires. Les performances sont comparées au modèle linéaire équivalent (5 :1).

Tableau 4-5 : Performances des modèles annuels pour la prévision de TURB_ATW

Modèle annuel - Répartition ATW 1		Toutes les données			Automne		Printemps			Eté	
		r	EQM	EAM	r	EAM	r	EQM	EAM	r	EAM
Test	Linéaire	0,86	0,322	**0,168**	**0,66**	**0,152**	**0,93**	**0,326**	**0,232**	0,79	**0,156**
	PMC 5:17:1	**0,87**	**0,302**	0,170	0,64	0,164	0,92	0,333	0,239	0,79	0,159
Toutes les données	Linéaire	0,90	0,393	**0,188**	**0,78**	**0,163**	0,91	0,665	**0,319**	**0,79**	**0,159**
	PMC 5:17:1	**0,91**	**0,367**	0,194	0,77	0,184	**0,93**	**0,590**	0,322	0,77	0,164

Modèle annuel - Répartition ATW 2		Toutes les données			Automne		Printemps			Eté	
		r	EQM	EAM	r	EAM	r	EQM	EAM	r	EAM
Test	Linéaire	0,81	0,548	0,199	0,78	**0,172**	0,75	1,13	0,359	**0,80**	**0,164**
	PMC 5:17:1	**0,87**	**0,480**	**0,176**	0,78	0,194	**0,84**	**0,923**	**0,346**	0,78	0,181
Toutes les données	Linéaire	0,90	0,393	**0,184**	0,77	**0,160**	0,91	0,676	**0,318**	**0,79**	**0,152**
	PMC 5:17:1	**0,91**	**0,374**	0,201	0,77	0,182	**0,93**	**0,612**	0,325	0,76	0,169

r : coefficient de corrélation

EQM : erreur quadratique moyenne (en UTN)

EAM : erreur absolue moyenne (en UTN)

Sur le Tableau 4-5, les modèles les plus performants sont indiqués en gras pour un ensemble et un critère de sélection fixé. En termes de corrélation, il n'y a pas de différences notables entre PMC et modèle linéaire. Une bonne corrélation sur l'ensemble Test est obtenue avec le modèle linéaire : de 0,81 à 0,86 sur toute l'année, et de 0,66 à 0,93 par saison. Concernant l'EAM, le modèle linéaire l'emporte plus

régulièrement, mais la différence est souvent infime (de l'ordre de quelques millièmes d'UTN). Quelques différences sur l'EQM pourraient favoriser les réseaux neuronaux : des améliorations de l'ordre de plusieurs dixièmes d'UTN sont observées au printemps et sur toute l'année. Les PMC pourraient avoir des prédictions plus précises pour les pointes de turbidité.

Cependant, vu que le critère principal (corrélation) ne présente pas de différence notable vis-à-vis des PMC et modèles linéaires, ce sont ces derniers qui seront conservés dans l'étape 2.

4.2.2 Étape 2 : modèles saisonniers

Un modèle linéaire par saison (soit trois modèles) est élaboré avec les entrées TURB_ATW-1 et TURB_DB-2. Ces performances sont comparées au modèle linéaire fonctionnant sur toute l'année (Tableau 4-6).

Tableau 4-6 : Performances des modèles saisonniers pour la prévision de TURB_ATW

Modèles Linéaires - Répartition ATW1	Toutes les données			Automne		Printemps			Eté	
	r	EQM	EAM	r	EAM	r	EQM	EAM	r	EAM
Test	0,86	0,322	0,168	0,66	0,152	0,93	**0,326**	0,232	0,79	0,156
				0,66	*0,152*	*0,93*	*0,327*	***0,226***	*0,79*	***0,150***
Toutes les données	0,90	0,393	0,188	0,78	**0,163**	0,91	0,665	0,319	0,79	0,159
				0,78	*0,169*	***0,93***	***0,599***	***0,308***	***0,80***	***0,149***

Modèles Linéaires - Répartition ATW2	Toutes les données			Automne		Printemps			Eté	
	r	EQM	EAM	r	EAM	r	EQM	EAM	r	EAM
Test	0,81	0,548	0,199	0,78	**0,172**	0,75	1,13	0,359	0,80	0,164
				0,84	*0,175*	***0,82***	***0,984***	***0,338***	***0,81***	***0,158***
Toutes les données	0,90	0,393	0,184	0,77	**0,160**	0,91	0,676	0,318	0,79	0,152
				0,78	*0,174*	***0,92***	***0,621***	***0,302***	***0,80***	***0,147***

Linéaire annuel

Linéaire saisonnier

Un gain d'un à sept dixièmes est obtenu pour la corrélation. L'EAM se trouve aussi améliorée de quelques centièmes, et les pointes de turbidité du printemps se trouvent

mieux prédits : les gains sur l'EQM montent jusqu'à 0,55 UTN. Le découpage en saison a ainsi permis d'améliorer légèrement les prédictions.

4.2.3 Étape 3 : identification des pointes mal prédits

Le tracé des graphiques de points éparpillés TURB_ATW prédite en fonction de TURB_ATW observée serait une droite d'équation y=x si le modèle était parfait. Une prédiction de pente inférieure à un est sous-estimée, si elle est au-dessus de cette droite elle est surestimée.

Au printemps

Pour les répartitions ATW 1 et 2, deux pointes pourraient présenter des problèmes au traitement, les autres sont prédites supérieures ou égales à 6 UTN. Le premier est le 13 mai 2000, prédit à 2 UTN au lieu des 6,3 UTN observés. Le deuxième est le 6 mars 2005, 4,8 UTN au lieu des 7,9 UTN observés (Figure 4-6).

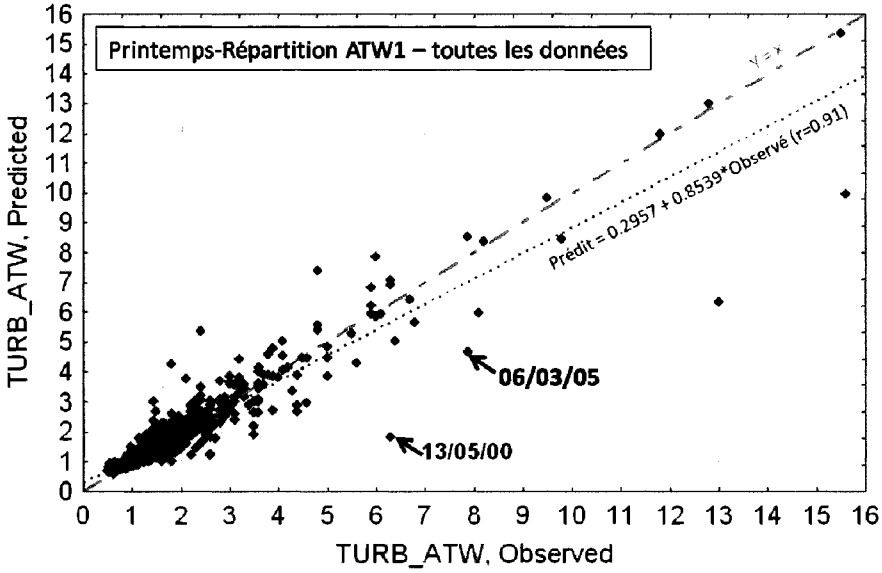

Figure 4-6 : TURB_ATW prédite en fonction d'observée-printemps-répartition ATW1-toutes les données

Le premier pic (celui du 13 mai 2000) n'est pas associé à la qualité de l'eau en amont ni même à la fonte des neiges. Les jours précédents, la turbidité de l'eau brute à Des Baillets reste en dessous de 5 UTN, et la contribution des Outaouais dans le fleuve Saint-Laurent reste inférieure à 10 %. Par contre, de fortes précipitations ont été enregistrées à Dorval quatre et trois jours avant (respectivement 29 et 21mm). Il pourrait s'agir de ruissellement urbain.
Pour le deuxième pic, celui du 6 mars 2005, ici aussi 27 mm de pluie ont été enregistrés à Dorval quatre jours avant.

Cependant, l'inclusion des variables DOR_PREC-4 ou IDX_FONT-4 n'apporte aucune amélioration notable sur ces deux évènements. Le modèle de l'étape 2 sera conservé.

À l'automne

Seulement deux pics sont moins bien prédits : le premier prédit à 4,2 UTN alors qu'observé 5,6 UTN, et le second prédit à 3 UTN alors qu'observé à 6,3 UTN.

Les causes de ces pointes sont déjà incluses en entrées (forte turbidité à Des Baillets deux jours avant), cependant la limitation à la bonne prédiction vient du manque d'exemples hauts. La grande proportion d'exemples de faible turbidité oriente préférentiellement le modèle vers leur apprentissage, au détriment des évènements de pointe.

Les modèles de régression et les entrées finalement retenus pour la prédiction de la turbidité à l'eau brute de l'usine Atwater sont récapitulés au Tableau 4-7.

Tableau 4-7 : Tableau récapitulatif des modèles retenus pour la prédiction de la turbidité à l'eau brute à Atwater

TURB_ATW	Saisonnier		
	Automne	Printemps	Été
Type	Linéaire	Linéaire	Linéaire
	2:1	2:1	2:1
Entrées	TURB_DB-2	TURB_DB-2	TURB_DB-2
	TURB_ATW-1	TURB_ATW-1	TURB_ATW-1
r TEST	0,84	0,93	0,81
EAM (TEST) en UTN	0,175	0,226	0,158

4.2.4 Étape 4 : modèle de classification

La section ci-dessus met en évidence que certaines pointes ne peuvent être bien classées par les modèles de régression retenus. L'implantation d'un modèle classifiant spécialisé sur ces pointes permettrait de prédire ces cas particuliers.

Cependant, en pratique pour cet ensemble de données le faible nombre d'exemples hauts nécessite d'élaborer un modèle de classification annuel. Sur toute l'année, ce dernier n'arriverait pas à séparer aisément les évènements identifiés ci-dessus.

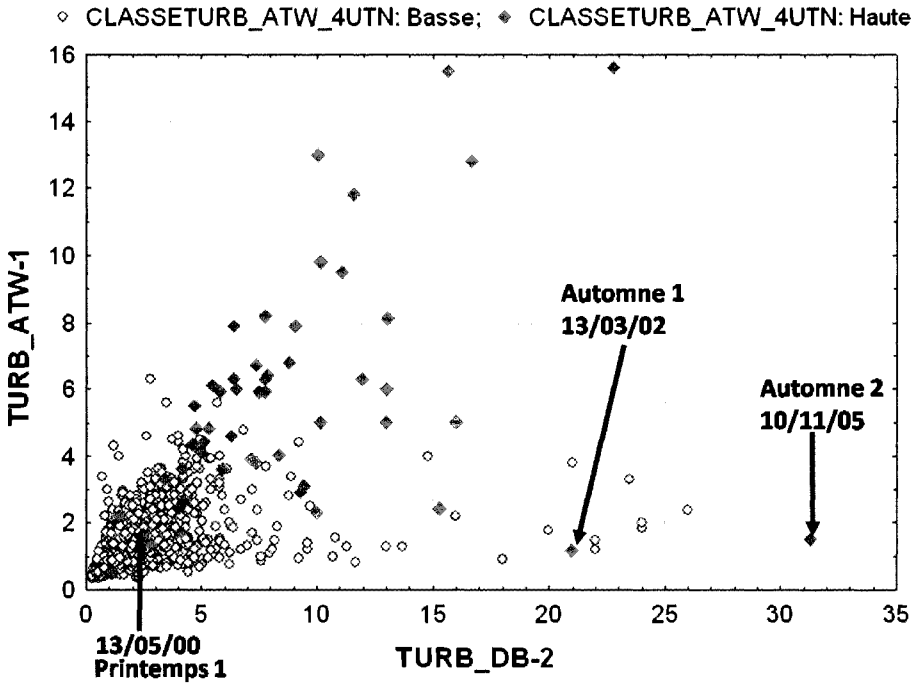

Figure 4-7 : Diagramme de points catégorisés - Classes de turbidité TURB_ATW-toutes les données

Sur la Figure 4-7, les trois évènements de pointe mal prédits par les modèles de régression présentent un fort chevauchement avec la classe basse (printemps 1 et automne 1) ou constituent un point isolé d'amplitude exceptionnelle (automne 2). Il semble évident que les descripteurs TURB_ATW-1 et TURB_DB-2 ne suffisent pas à séparer ces évènements exceptionnels. Un modèle de classification annuel basé sur ces descripteurs ne permettrait donc pas d'obtenir une performance additionnelle.

4.3 Discussion

Quelques remarques générales sur les modèles développés sont présentées.

Remarque sur les modèles linéaires

Tout d'abord, concernant les répartitions utilisées, l'utilisation d'un modèle linéaire ne requiert pas de partitionner les données en trois ensembles : en effet, les modèles

linéaires ne sont pas sujets au sur-apprentissage. L'ensemble Select est donc superflu, il pourrait être incorporé à l'ensemble d'apprentissage pour augmenter la précision de prédiction. Il a cependant été conservé tel quel afin de pouvoir comparer les modèles linéaires au modèle PMC annuel.

Les modèles linéaires développés ici ne comprennent que les composantes du premier ordre. Un modèle régressif plus poussé pourrait inclure les composantes du deuxième ordre, ainsi que les effets d'interactions entre les entrées.

La dernière remarque porte sur l'usage d'un modèle linéaire et sur les corrélations. Au printemps, les deux entrées TURB_ATW-1 et TURB_DB-2 sont fortement corrélées (r=0,83, Tableau 4-2). De même que ces entrées sont corrélées avec la sortie TURB_ATW. Contrairement aux modèles RNA où cela ne semble pas poser problème, il est d'usage courant en modélisation linéaire de différencier afin que les séries temporelles soient stationnaires. Voir à ce sujet la méthode de Haugh & Box (Maier et Dandy, 1997). Une telle opération n'a pas été effectuée ici, toutefois, des prédictions suffisamment bonnes ont pu être obtenues.

Remarque sur l'utilisation potentielle des RNA

La piste de l'implantation d'un modèle neuronal au lieu d'un modèle linéaire serait à envisager. Bien que les performances des modèles linéaires présentés soient convenables, le modèle PMC annuel semble montrer une légère amélioration sur un des pointes problématique du printemps (données encerclées de la Figure 4-8). Le modèle PMC semble moins sous estimer la prédiction que les modèles linéaires. Cependant, le modèle linéaire saisonnier obtient des prédictions semblables. Un modèle neuronal saisonnier pourrait améliorer d'avantage ces prédictions.

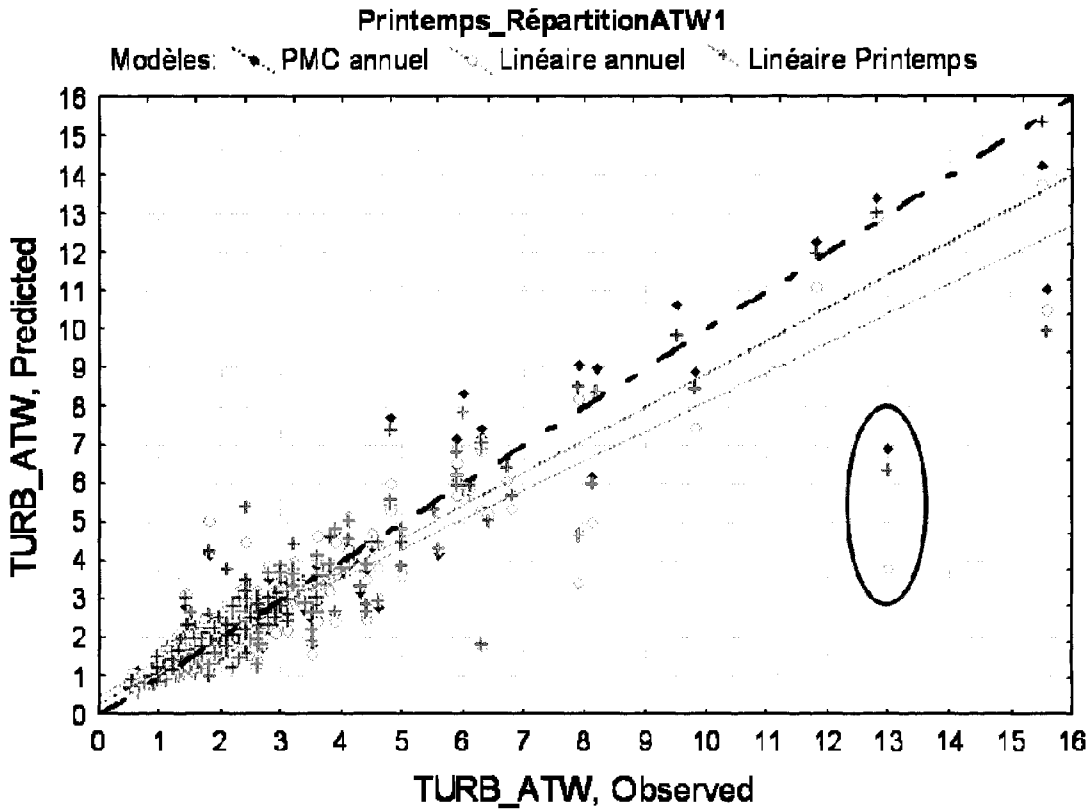

Figure 4-8 : Diagramme de points éparpillés TURB_ATW observée en fonction de prédite pour les trois modèles - Printemps – Répartition_ATW1

Chapitre 5 PRÉDICTION DE LA TURBIDITÉ À L'EAU FILTRÉE À LA STATION DES BAILLETS

5.1 Étapes de la modélisation

5.1.1 Objectifs et mise en contexte

L'usine de Des Baillets fonctionne actuellement avec des filtres à sable non assistés chimiquement (sans ajout de coagulant). Elle comporte 60 filtres ayant 160m^2 de surface chacun, et ils sont répartis en quatre galeries de quinze filtres. Le milieu filtrant est constitué de sable de granulométrie uniforme (0,6mm) sur 1,2m de hauteur. L'usine opère la plupart du temps à une charge superficielle relativement constante de 5 m/h.

L'objectif de cette partie est de bâtir un modèle prédictif de la turbidité en sortie des filtres une journée à l'avance (variable appelée TURF_DB).

Afin de rencontrer les normes du RQEP (2005), les filtres doivent produire une turbidité à l'eau filtrée sous une valeur seuil, soit 1 UTN en moyenne mobile mensuelle pour la filtration non-assistée chimiquement.

En cas de dégradation rapide de la qualité de l'eau brute, ce modèle permettrait de prédire un éventuel dépassement des normes, et ainsi de mettre en place des moyens d'action préventifs.

Nous émettons comme hypothèse que la turbidité à l'eau filtrée est sous l'influence directe de la qualité de l'eau brute (turbidité, couleur, etc.), de la turbidité à l'eau filtrée aux pas précédents (ceci reflétant sa capacité à être traitée), des propriétés physiques de l'eau (température, viscosité, etc.), et de la saison (type de particules).

5.1.2 Base de données disponible

Données disponibles

Les variables disponibles avec leurs résolutions temporelles, et les dates auxquelles elles sont disponibles figurent au Tableau 5-1.

Tableau 5-1 : Données physico-chimiques disponibles à la station Des Baillets

Variables	Point de mesure	Résolution	Date de début	Date de fin	Détail
Température	Brute	aux 8 jours environ	01-Jan-96	31-Dec-06	
	Traitée	journalier (avec trous)	01-Jan-96	31-Dec-06	
Couleur	Brute	journalier	01-Jan-96	31-Dec-06	UCA
		hebdomadaire	01-Jan-96	31-Dec-06	UCV
	Filtrée	journalier	01-Jan-96	31-Dec-06	UCA : filtres N-O, S-O, N-E, S-E
		hebdomadaire	01-Jan-96	31-Dec-06	UCV : mixte
Turbidité	Brute	journalier	01-Jan-96	31-Dec-06	UTN
		4 heures	01-Jan-00	14-Aug-05	
	Filtrée	journalier	01-Jan-96	31-Dec-06	UTN : filtres N-O, S-O, N-E, S-E
		4 heures	01-Jan-00	01-Apr-07	
		4 heures	01-Jan-01	01-Apr-07	R46113, R46123, R46213, R46223
		2 heures	15-Aug-04	30-Jun-06	Filtres 1 à 60
Conductivité	Traitée	journalier	01-Jan-96	31-Dec-06	
Alcalinité	Brute	hebdomadaire	01-Jan-96	31-Dec-06	
	Traitée	hebdomadaire	01-Jan-96	31-Dec-06	
Dureté totale	Brute	hebdomadaire	01-Jan-96	31-Dec-06	
	Traitée	hebdomadaire	01-Jan-96	31-Dec-06	
Carbone organique dissous	Brute	journalier	01-Jan-96	31-Dec-06	
	Traitée	journalier	01-Jan-96	31-Dec-06	
pH	Brute	journalier	01-Jan-96	31-Dec-06	
	Ozonée	journalier	01-Jan-96	31-Dec-06	
	Traitée	journalier	01-Jan-96	31-Dec-06	

Contrairement aux articles cités lors de la revue de littérature (Section 1.3.3), il n'y a pas d'ajout de coagulant ici. Ainsi, les variables telles que le pH, l'alcalinité, la dureté totale, etc. ne devraient pas influencer significativement la performance de la filtration.

Pour ce qui est de la température de l'eau brute (TEMPEB_DB), celle-ci peut affecter les propriétés physiques de l'eau, dont la viscosité. Comme elle pourrait être un

paramètre important, il faut combler les données manquantes. Ceci est effectué par interpolation linéaire. Cette variable étant relativement constante dans le temps, il ne sera pas utile d'explorer les décalages temporaux de la température de l'eau brute.

La conductivité varie peu, une meilleure information de la qualité de l'eau (et des rapports de mélange des masses d'eau) serait contenue dans la turbidité à l'eau brute. Cette variable sera donc écartée.

La couleur à l'eau brute varie peu, mais elle peut être représentative d'une eau plus chargée en particules colloïdales, donc plus difficile à traiter par filtration sur sable sans coagulation.

La vitesse de filtration ou le débit journalier produit par l'usine n'ont pas été disponibles au moment de l'élaboration des modèles. Ces données n'ont pu être obtenues avant le début des analyses. Leurs valeurs sont importantes, elles contribueraient sans aucuns doutes à améliorer la performance des modèles conçus ici.

Pour la turbidité à l'eau filtrée, des données sont disponibles sur divers pas de temps :

- Du 1er janvier 96 au 31 mai 2006 en moyenne journalière par galerie (nord ouest, nord est, sud ouest, et sud est).

- Du 1er janvier 2000 au 14 aout 2005 en moyenne aux 4 heures.

Les valeurs enregistrées pour les quatre galeries sont assez semblables. Nous émettons l'hypothèse que ces dernières sont identiques. Une variable additionnelle est créée, c'est la moyenne de ces 4 galeries. Elle se nomme TURF_DB.

Quel pas de temps choisir pour la turbidité ?

Il est important de se fixer un pas de temps pour l'analyse de la turbidité. Ce pas de temps est dépendant des données disponibles, mais surtout de la vitesse de variation

du phénomène à modéliser. Si ce dernier varie vite, des prévisions à court terme permettront de capturer ces brusques variations. Au contraire, une variation lente du phénomène pourra se contenter d'un pas de temps plus long.

Des données d'eau filtrées mixte sont disponibles aux 4 heures de 2000 à 2007. Travailler sur ces données écarterait les fortes pointes de turbidité observés en 1998. Il faut vérifier que la turbidité à l'eau filtrée ne varie pas trop vite.

Pour ce faire, il convient d'analyser les statistiques descriptives de la valeur absolue de la variation de la turbidité à l'eau filtrée : Delta TURF_DB t- x heures. La variable x prend les valeurs 4, 8 et 16 heures (Tableau 5-2).

Tableau 5-2 : Statistiques descriptives des variations de turbidité à l'eau filtrée de l'usine Des Baillets

	N	Moyenne	Min	Max	P25	P75	P95	P99	Ecart type
Delta TURF_DB_4H	12280	0,02	0	1,11	0	0,02	0,09	0,25	0,06
Delta TURF_DB_8H	12274	0,03	0	1,61	0	0,03	0,11	0,32	0,07
Delta TURF_DB_16H	12267	0,03	0	1,54	0	0,03	0,12	0,4	0,08

Les moyennes et les 75ème centiles observés sont très faibles (de l'ordre de quelques centièmes d'UTN). Même le 99ème centile est seulement de 0,4 UTN en 16 heures d'écart.

En conclusion, les variations observées sur la turbidité à l'eau filtrées sont faibles. TURF_DB variant lentement, il est possible de se contenter des données journalières moyennes. Le modèle devra prédire une journée à l'avance la valeur moyenne quotidienne de la variable TURF_DB (elle-même moyenne des quatre galeries de filtres).

5.1.3 Inspection graphique de TURF_DB

Le tracé de la turbidité à l'eau filtrée en fonction de la date julienne est donné à la Figure 5-1. Les découpages saisonniers utilisés pour la prédiction de la turbidité à l'eau brute pour Atwater et Des Baillets figurent encore sur le graphe.

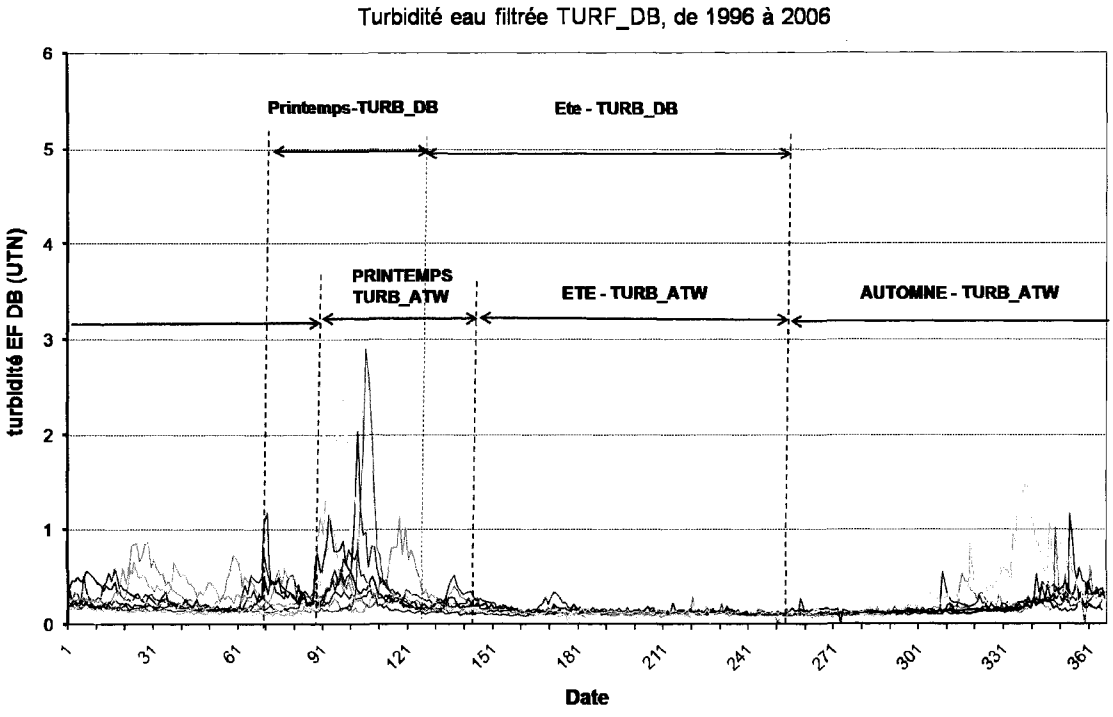

Figure 5-1: TURF_DB en fonction de la date julienne de 1996 à 2006

De nouveau trois saisons se distinguent. Le découpage en saison utilisé pour Des Baillets correspond bien à l'exception du printemps, où le début serait le même que le printemps à Atwater. La fin serait encore celle du printemps à Des Baillets. Afin de calculer les performances de prédiction par station, les saisons de Des Baillets seront conservées. Ceci ne présente pas d'inconvénient majeur dans la mesure où la saison de Des Baillets englobe tous les pointes printanières.

Concernant les deux périodes principales, automne et printemps, la plus grande intensité est obtenue au printemps (5,2 UTN le 4 avril 1998), alors que l'automne

monte jusqu'à un maximum de 1,51 UTN (le 5 décembre 2003). Les deux saisons présentent des pointes de durée prolongée.

À l'automne, une dégradation générale de la qualité de l'eau filtrée est observée, mais les valeurs sont quasiment inférieures à 1UTN. Tout le reste du temps (été y compris), la TURF_DB reste négligeable (<0,5UTN).

Ici aussi, il y a un inversement de tendance par rapport à l'eau brute : les pics de turbidité printaniers sont plus difficiles à traiter, bien qu'ils soient de moindre amplitude par rapport aux événements automnaux. Ce comportement a déjà été observé après le passage dans le canal Atwater. Ceci renforce l'hypothèse que les particules responsables de la turbidité à l'automne ne sont pas les mêmes qu'aux printemps; elles seraient plus grosses pour pouvoir être décantées ou enlevées sur les filtres à sable. Cet inversement de tendance impose aussi l'élaboration de modèles saisonniers : ce ne sont pas les mêmes phénomènes à modéliser au printemps et à l'automne.

5.1.4 Analyse statistique de TURF_DB

Les statistiques descriptives sont données au Tableau 5-3 pour toutes les données, et pour chaque saison.

Tableau 5-3 : Statistiques descriptives de TURF_DB, pour l'année et par saisons

	TURF_DB (en UTN)									
	N	Moyenne	Médiane	Minimum	Maximum	P25	P75	P90	P95	Écart type
Toutes les données	3769	0,22	0,15	0,04	5,20	0,11	0,24	0,38	0,52	0,27
Printemps	616	0,43	0,28	0,10	5,20	0,19	0,44	0,80	1,30	0,55
Été	1333	0,13	0,12	0,04	0,52	0,10	0,14	0,18	0,22	0,05
Automne	1322	0,21	0,17	0,05	1,51	0,12	0,26	0,38	0,47	0,16

Sur toute l'année

Sur toute l'année (3769 exemples au total), la plupart des valeurs sont comprises entre 0,04 et 0,15 UTN (médiane). Un maximum de 5,2 UTN est obtenu au printemps 1998. Les données sont relativement concentrées autour de faibles valeurs de turbidité : le 95ème centile reste inférieur à 0,52 UTN (et le 98ème centile est de 0,85 UTN). Moins de 2% des mesures dépassent le seuil de 1 UTN. Sur les dix ans, seulement une observation supérieure à 5 UTN a été observée.

Par saison

La qualité reste excellente dans le détail des saisons : dans tous les cas, le 90ème centile est inférieur ou égal à 0,80 UTN, il est inférieur à 0,5 UTN pour l'automne et l'été. La qualité à l'été est excellente (moyenne de 0,13 et écart-type de 0,05 UTN), ensuite vient l'automne (0,21 et 0,16 UTN respectivement). Seul le printemps présente un risque avec un peu moins de 10% des données dépassant la norme de 1 UTN.

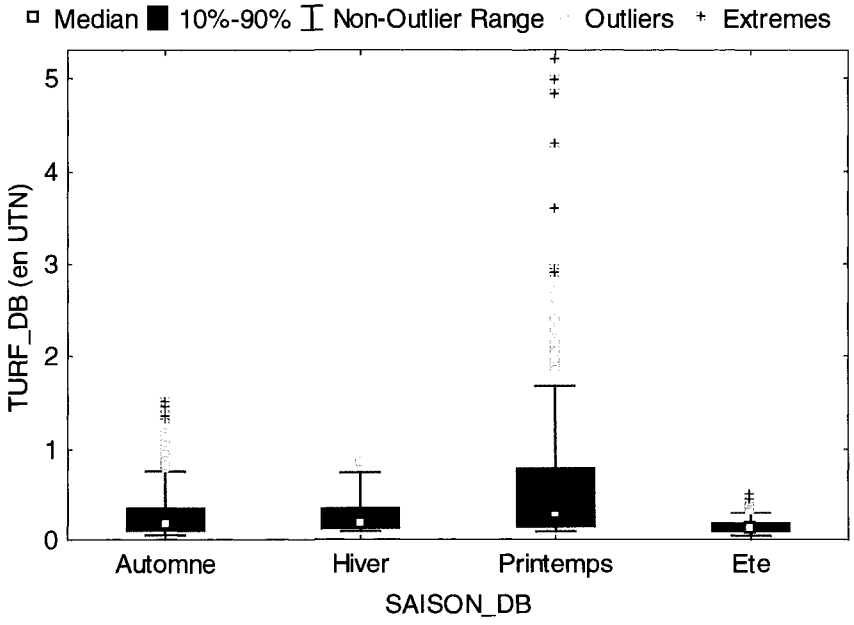

Figure 5-2 : Variations saisonnières de la turbidité à l'eau filtrée TURF_DB

5.1.5 Partitionnement des exemples

Afin de pouvoir joindre la prédiction de l'eau filtrée à celle de l'eau brute dans une même interface prédictive, le partitionnement des exemples utilisé précédemment (Chapitre 3) sera de nouveau utilisé ici. Il comprend deux répartitions (DB 98 et 99) issues d'un échantillonnage aléatoire à proportion fixée parmi des classes de turbidité.

5.1.6 Choix des entrées

Analyse des corrélations

Tableau 5-4 : Corrélations croisées de TURF_DB avec les variables de qualité de l'eau

	TURF_DB			
	Toute l'année	Automne	Été	Printemps
TEMPEB_DB-1	-0,32	-0,48	-0,50	-0,03
COUL_DB-1	0,59	0,66	0,72	0,67
TURB_DB	0,55	0,46	0,25	0,81
TURB_DB-1	0,51	0,35	0,30	0,78
TURF_DB-1	0,93	0,85	0,87	0,93

Quelle que soit la saison, il y a une forte corrélation entre TURF_DB et sa valeur la veille (r compris entre 0.85 et 0.93). À l'automne et à l'été, la couleur la veille (COUL_DB-1) est la deuxième variable la plus corrélée avec la turbidité à l'effluent, alors qu'au printemps, la turbidité revêt plus d'importance.

Pour le modèle linéaire annuel, TURF_DB-1 est fortement corrélée avec la valeur du lendemain : il sera inclus d'office dans tous les modèles. Pour les autres variables, elles semblent d'égale importance en terme de corrélation. Or, a posteriori, quelques essais sur les différentes combinaisons entre TURF_DB-1, TEMPEB_DB-1, TURB_DB-1 et COUL_DB-1 révèlent que l'inclusion de deux ou quatre variables ne change pas les performances obtenues (en termes de corrélation et d'EAM). Ainsi,

seules les variables TEMPEB_DB-1 et TURF_DB-1 seront conservées. La variable de température véhicule l'information sur la saison.

Pour les modèles saisonniers, les quatre variables TURF_DB-1, TURB_DB-1, COUL_DB-1, et TEMPEB_DB-1 semblent fortement corrélées avec la sortie. Elles seront donc les candidates pour les modèles saisonniers. Pour l'été, la turbidité à l'eau brute est relativement faible, la couleur permettrait de déceler les pointes de matières dissoutes à l'eau brute, matière difficile à traiter par les filtres à sable.

5.1.7 Tableau récapitulatif des modèles retenus

Les entrées des modèles et l'architecture de ces derniers ont été déterminées par l'algorithme IPS de Statistica. Les paramètres internes utilisés sont les mêmes que pour le modèle régressif du Chapitre 3. Les modèles finaux retenus sont décrits dans le Tableau 5-5.

Tableau 5-5 : Résumé des modèles prédictif retenus pour la turbidité à l'eau filtrée - TURF_DB

	Annuel	Saisonnier		
		Automne	Printemps	Été
Type	Linéaire	PMC	PMC	Linéaire
	2:1	5:7:1	4:10:1	2:1
Entrées	TURF_DB-1	TURF_DB-1	TURF_DB-1	TURF_DB-1
	TEMPEB_DB-1	TEMPEB_DB-1	TEMPEB_DB-1	COUL_DB-1
		COUL_DB-1	COUL_DB-1	
		TURB_DB-1	TURB_DB-1	
		LSF_VITM-1		

Il ressort que l'automne et le printemps ont conservé les quatre variables initialement candidates. Pour l'été, les variables TURF_DB-1 et COUL_DB-1 sont retenues. De plus, vu qu'aucun pic supérieur à 0,8 UTN n'a été observé l'été, un modèle linéaire est suffisant pour obtenir un coefficient de corrélation supérieur à 0,88 (voir Section 5.2).

Pour l'automne, quelques modèles élaborés par essais et erreurs (modèles non présentés ici) ont montrés que les données de la veille sont insuffisantes pour prédire un pic de turbidité le lendemain. Ainsi une autre variable explicative des pointes de turbidité à l'eau brute (LSF_VITM-1) fut incluse afin d'améliorer les prédictions à l'eau filtrée.

5.2 Résultats

Sur le Tableau 5-6, sont affichés les résultats des modèles pour les critères suivants : le coefficient de corrélation (r), l'erreur quadratique moyenne (EQM en UTN), et l'erreur absolue moyenne (EAM en UTN); ceci selon les deux répartitions utilisées, sur les ensembles de test, sur toutes les données, et sur toutes les données où TURF_DB a été observée supérieure ou égale à 0,8 UTN. Ce dernier ensemble permet de représenter la performance de prédiction pour les valeurs critiques de turbidité en sortie des filtres. Les performances du modèle linéaire sont comparées à celles des modèles saisonniers (en italique sur la deuxième ligne), le plus performant des deux est inscrit en gras.

Les résultats montrent que le modèle saisonnier d'été apporte une bonne amélioration de la prédiction sur tous les ensembles de données. Aucun dépassement supérieur à 0,8 UTN n'est observé, c'est pourquoi la rubrique est vide.

En ce qui concerne le printemps, les résultats sont assez variables : les différences entre modèles saisonniers et linéaire annuel sont de l'ordre de quelques centièmes pour les ensembles Test et toutes les données. Le modèle le plus performant (linéaire annuel ou PMC saisonnier) varie selon la répartition considérée. Par contre, pour les données critiques (les observations supérieures à 0,8 UTN), le modèle printanier semble diminuer l'EAM dans les deux répartitions.

Tableau 5-6 : Comparaison des performances des modèles annuels et saisonniers pour la prévision de TURF_DB

Répartition DB 99	Toutes les données			Automne		Printemps			Été	
	r	EQM	EAM	r	EAM	r	EQM	EAM	r	EAM
Test	0,908	0,077	0,031	**0,929**	**0,035**	**0,956**	0,22	0,11	0,875	0,017
				0,903	*0,038*	*0,889*	*0,16*	***0,080***	***0,890***	***0,015***
Toutes les données	0,929	0,099	0,040	0,851	0,040	0,929	0,20	**0,094**	0,877	0,017
				0,878	*0,040*	***0,931***	*0,20*	*0,099*	***0,883***	***0,016***
TURF_DB obs >= 0,8 UTN	0,861	0,54	0,41	0,519	0,33	0,857	**0,58**	0,49	Aucun élément observé >0,8	
				0,665	***0,32***	***0,865***	*0,60*	***0,45***		
	Linéaire 2:1			*PMC 5:7:1*		*PMC 4:10:1*			*Linéaire 2:1*	

Répartition DB 98	Toutes les données			Automne		Printemps			Été	
	r	EQM	EAM	r	EAM	r	EQM	EAM	r	EAM
Test	0,910	0,17	0,052	0,773	0,043	0,912	0,22	**0,071**	0,795	0,020
				0,875	***0,040***	***0,914***	***0,17***	*0,15*	***0,901***	***0,019***
Toutes les données	0,929	0,099	0,040	0,851	0,040	0,929	0,20	**0,094**	0,877	0,017
				0,880	*0,040*	***0,931***	*0,20*	*0,096*	***0,883***	***0,016***
TURF_DB obs >= 0,8 UTN	0,861	0,55	0,41	0,518	0,34	**0,857**	0,60	0,45	Aucun élément observé >0,8	
				0,602	***0,31***	*0,846*	***0,56***	***0,43***		
	Linéaire 2:1			*PMC 5:7:1*		*PMC 4:10:1*			*Linéaire 2:1*	

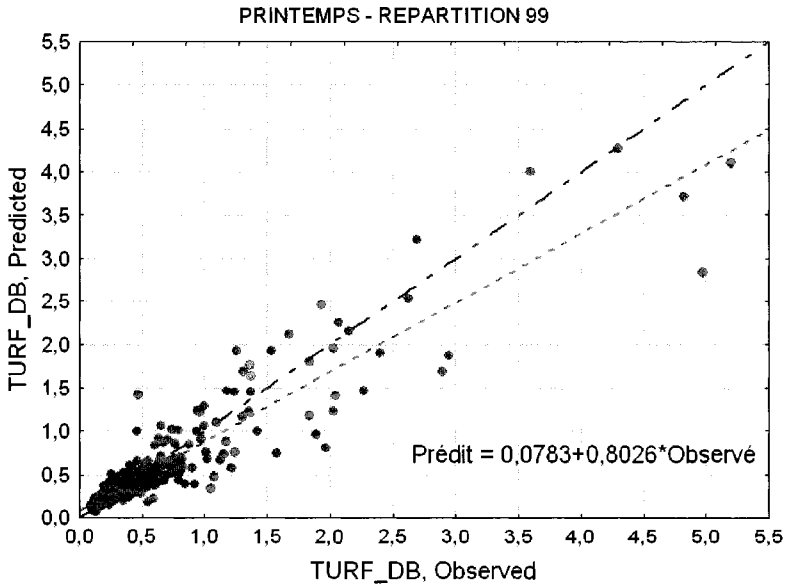

Figure 5-3 : Turbidité à l'eau filtrée de Des Baillets - TURF_DB prédite en fonction de la turbidité observée; printemps-répartitionDB99-toutes les données (n=616)

Sur la Figure 5-3, sont tracées les prédictions données par le modèle en fonction des valeurs observées. Une bonne concordance à basse turbidité est observée, mais au-delà de 1 UTN, le modèle a tendance à sous-estimer les prédictions : neuf valeurs observées supérieures à 1 UTN (sur un total de 45) sont prédites en dessous de 0,8 UTN, soit 20 % de mauvaise classification. Toutefois, sur ces neuf valeurs mal classées, seulement trois ont été prédites en dessous de 0,6 UTN.

Concernant les résultats de l'automne, le Tableau 5-6 montre une légère amélioration du PMC sur toutes les données d'automne pour les deux répartitions. En effet, la corrélation passe d'environ 0,85 à 0,88 dans les deux répartitions. En revanche, les améliorations sont notables pour les évènements supérieurs à 0,8 UTN. Par exemple, pour la répartition 99, la corrélation passe de 0,519 à 0,665 avec le modèle PMC 5 :7 :1. La Figure 5-4 montre que les résultats prédits sous-estiment encore les valeurs observées, et que trois valeurs sur les neuf observées supérieures à 1 UTN sont prédites en dessous de 0,8 UTN.

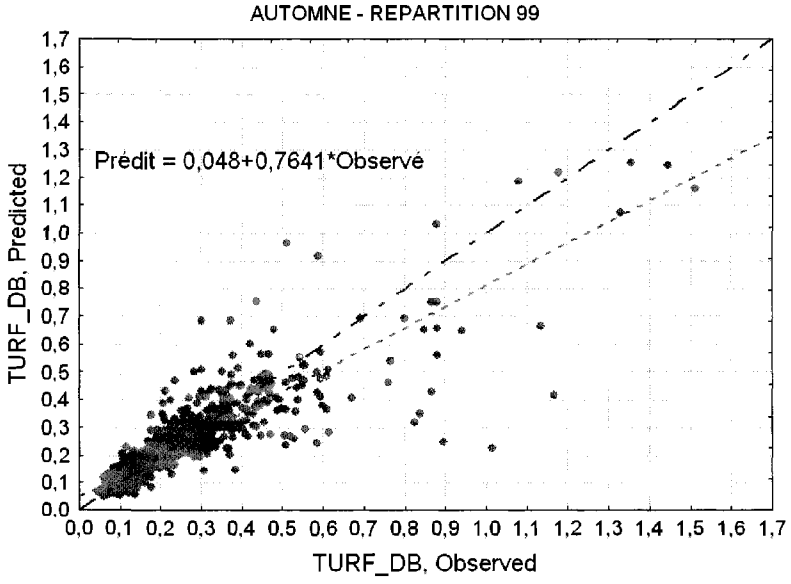

Figure 5-4 : TURF_DB prédite en fonction d'observée-automne-répartitionDB99-toutes les données (n=1322)

5.3 Discussion

5.3.1 Besoin d'un modèle de classification ?

De gros écarts sont observés entre les pointes extrêmes et le reste des données ($10^{ème}$ à $90^{ème}$ centile sur le Tableau 5-3). Un modèle de classification pourrait être utile afin de se spécialiser dans la prédiction de ces pointes. Il pourrait ainsi améliorer la performance de prédiction du modèle régressif.

L'automne présente 21 exemples supérieurs ou égaux à 0,8 UTN, alors que le printemps en présente 61.

Le seuil 0,8 UTN peut présenter une valeur critique pour l'opération : la norme en sortie des filtres est de 1 UTN. De plus, le Chapitre 3 a révélé que les valeurs au voisinage du seuil de classification sont sujettes à de mauvaises classifications dû au chevauchement des classes. Le seuil 0,8 UTN semble être un bon compromis entre disposer de suffisamment de données supérieures à ce seuil, et présenter une bonne alarme pour l'opération de l'usine.

Sur la Figure 5-5, le diagramme de points catégorisés montre qu'une bonne séparation peut être obtenue pour le printemps (Fig. a), alors que l'automne présente plus de chevauchement des classes de turbidité (Fig. b). Les descripteurs utilisés sont la température et la turbidité à l'eau brute le même jour. Intuitivement, ces descripteurs semblent être les plus aptes à modéliser la turbidité en sortie des filtres. Ceci est vrai au printemps, où la majeure partie des évènements supérieurs à 0,8 UTN sont situés entre 1 et 6 °C. À l'automne, par contre, les filtres retiennent plusieurs pointes de turbidité à l'eau brute même si la température est inférieure à 6°C. Il faudrait donc trouver d'autres descripteurs comme les vitesses moyennes du vent sur les lac Saint-Louis et Saint-François. Par exemple, sur la Figure 5-6, les classes de turbidité par rapport au seuil 0,8 UTN sont tracées en fonction de LSF_VITM-1 et COUL_DB. La

couleur à Des Baillets le jour même s'avère être un bon descripteur, moins de chevauchement des classes est observé.

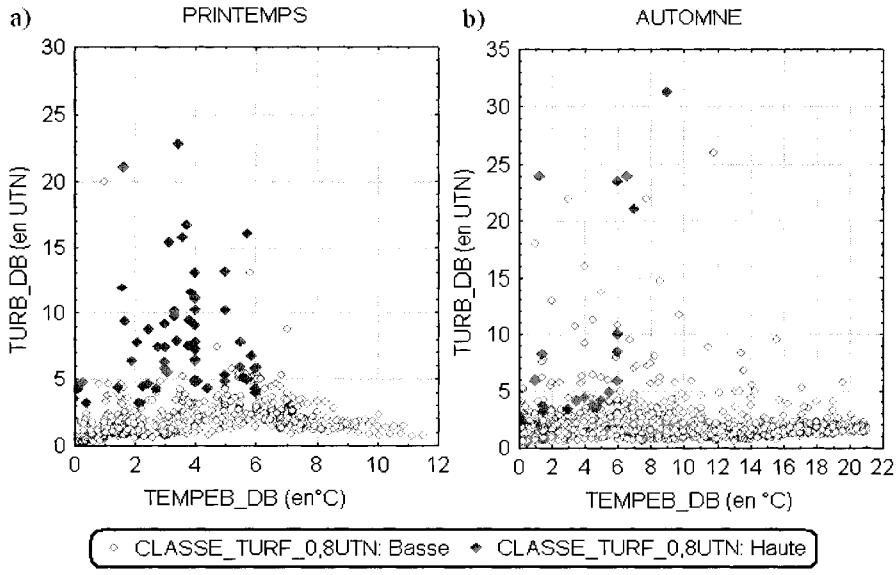

Figure 5-5 : Diagramme de points catégorisés de TURF_DB en fonction de TURB_DB et TEMPEB_DB, (a) au printemps, (b) à l'automne

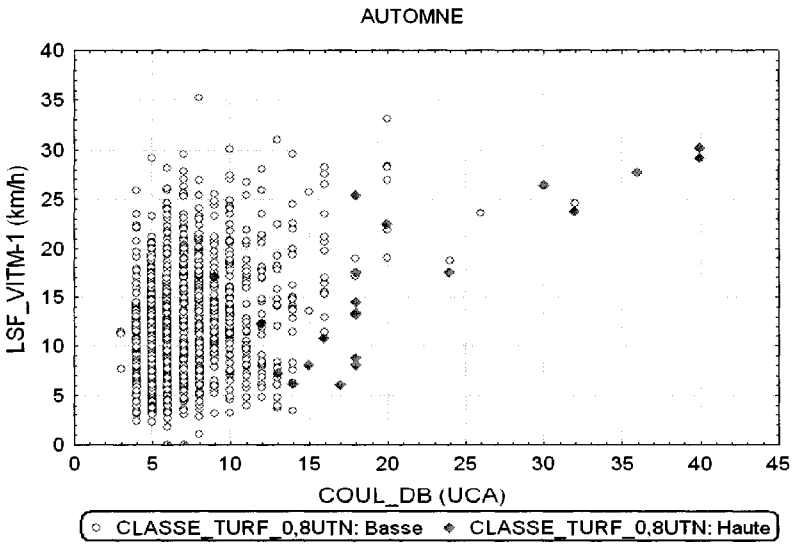

Figure 5-6 : Diagramme de points catégorisés de TURF_DB en fonction de LSF_VITM-1 et COUL_DB à l'automne

5.3.2 Améliorations des prédictions à l'automne et au printemps

Les résultats obtenus ici ne sont basés que sur les données de la qualité à l'eau brute, et de la turbidité à l'eau filtrée. Au printemps, une bonne corrélation existe entre TURF_DB et ces variables, ainsi des performances acceptables sont facilement accessibles (r de l'ordre de 0,88). Une première piste pour améliorer les prédictions pourrait tenir compte des causes explicatives de la turbidité à l'eau brute, et inclure ces causes en tant que variables d'entrées. Ceci à déjà été mené pour la saison automne avec l'inclusion de la vitesse moyenne du vent sur le lac Saint-François. Le coefficient de corrélation pour les données supérieures à 0,8 UTN a gagné +0,1 unité.

Une deuxième amélioration serait l'élaboration d'un modèle tenant compte des charges superficielles sur les filtres ou bien du débit d'eau journalier produit par l'usine. Ces valeurs affectent directement la qualité de l'eau en sortie des filtres. La notion de production d'eau est déjà 'incluse' dans la notion de saison : ce n'est pas le même débit qui est produit à l'été et au printemps. Cependant, cette notion pourrait être raffinée.

Une troisième amélioration possible serait la prise en compte des valeurs du lendemain en tant qu'entrées. L'étude des corrélations (Tableau 5-4) montre qu'il existe un lien plus fort entre la turbidité à l'eau brute et à l'eau filtrée le même jour (i.e. entre TURB_DB et TURB_ATW) qu'entre la turbidité à l'eau brute la veille et à l'eau filtrée le lendemain (i.e. entre TURB_DB-1 et TURB_ATW). Il doit en être de même avec la couleur. Il serait intéressant de bâtir un modèle utilisant en entrée les résultats de la prédiction donnée par le modèle prédictif de la turbidité à l'eau brute; i.e., prédisant TURF_DB à partir des valeurs prédites pour TURB_DB (ou bien COUL_DB, si un modèle prédictif de la couleur devait être développé). Reste à déterminer si le gain de performance obtenu par l'inclusion d'une entrée véhiculant plus d'information (par exemple TURB_DB) n'est pas compensé par l'erreur de prédiction commise par le modèle prédictif à l'eau brute.

Chapitre 6 PRÉDICTION DE LA TURBIDITÉ À L'EAU FILTRÉE À LA STATION ATWATER

6.1 Étapes de la modélisation

6.1.1 Objectifs et mise en contexte

Comme Des Baillets, l'usine Atwater fonctionne actuellement avec des filtres à sable non assistés chimiquement (sans ajout de coagulant). L'objectif de cette partie est de bâtir un modèle prédictif de la turbidité mixte en sortie des filtres une journée à l'avance (variable appelée TURF_ATW).

Le cheminement et la méthode utilisée sont les mêmes que pour le chapitre précédent.

L'usine possède 7 galeries de 16 filtres, soit un total de 112 filtres de superficie 111,5 m^2. L'épaisseur du milieu filtrant est 0,76 m; il s'agit de sable de granulométrie uniforme de 0,6 mm. La charge superficielle varie de 2,5 à 5 m/h. Les débits filtrés journaliers moyens ne sont disponibles qu'à compter de 2002.

6.1.2 Base de données disponible

Données disponibles

Les variables disponibles avec leurs résolutions temporelles, et les dates auxquelles elles sont disponibles sont sensiblement les mêmes que pour l'usine Des Baillets. Des données de turbidité ainsi que les débits de filtration par galeries furent disponibles aux deux heures à partir du 1er janvier 2002. Cependant cette date obligerait à négliger les années 1996 à 2001. Dans un premier temps, le modèle sera développé sans utiliser l'information contenue dans les débits de filtration.

Basé sur les résultats du chapitre précédent, les variables suivantes seront utilisées comme candidates potentielles pour l'élaboration du modèle : la qualité de l'eau brute à Des Baillets deux jours avant (turbidité, couleur, température de l'eau brute), à Atwater la veille, et la turbidité à l'eau filtrée d'Atwater la veille. De plus, si cela améliore les prédictions les variables suivantes seront aussi envisagées : OUT_FLV-3,

RIV_CHAT-5, IDX-FONT-3, LSF_VITM-3, et DOR_PREC-3-4. Elles correspondent aux entrées importantes des modèles saisonniers retenus pour la prédiction de TURB_ATW.

Le pas de temps adopté sera aussi journalier. La variable de sortie du modèle sera appelée TURF_ATW.

Plusieurs valeurs de turbidité à l'eau filtrée sont manquantes de 1996 à 1999. Les pics de turbidité printaniers de 1998 sont incomplets.

6.1.3 Inspection graphique de TURF_DB

Le tracé de la turbidité à l'eau filtrée en fonction de la date julienne est donné à la Figure 6-1. Les découpages saisonniers utilisés pour la prédiction de la turbidité à l'eau brute pour Atwater et Des Baillets figurent encore sur cette figure.

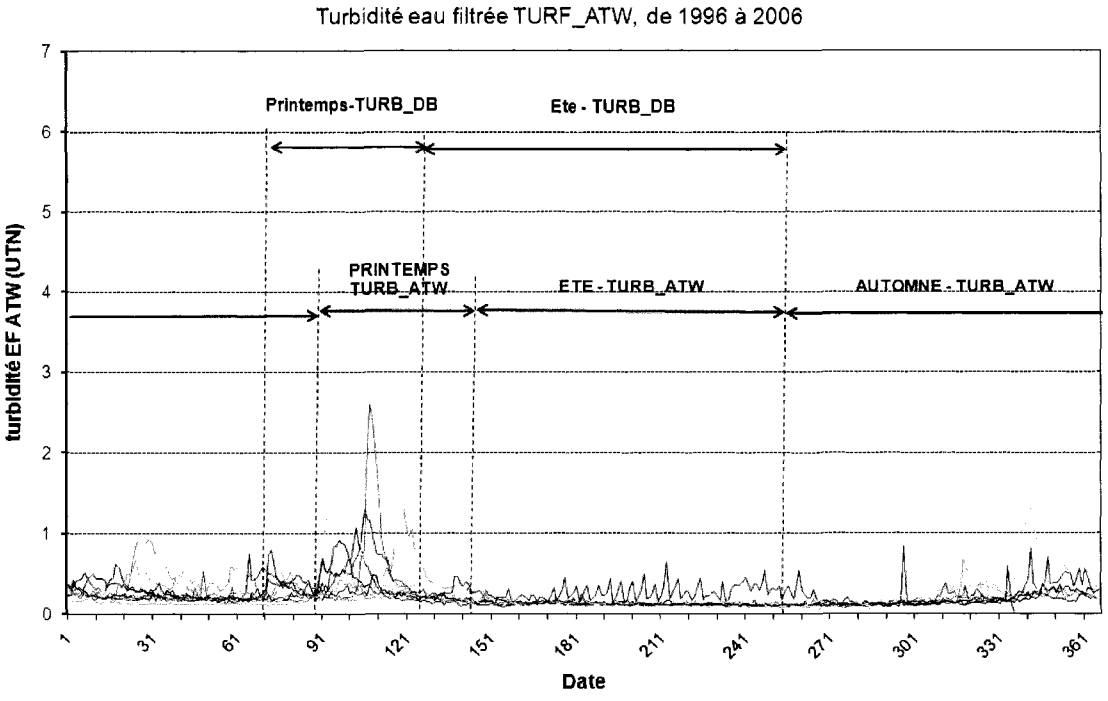

Figure 6-1 : TURF_ATW en fonction de la date julienne de 1996 à 2006

À nouveau, trois saisons se distinguent. Les caractéristiques des saisons sont identiques à celles identifiées au Chapitre 4. Des modèles saisonniers seront de

nouveau développés. Une bonne correspondance des saisons est observée avec les saisons identifiées pour le modèle de prédiction de TURB_ATW. Afin de calculer les performances de prédiction par station, les saisons d'Atwater seront conservées.
Concernant les deux périodes principales, l'automne et le printemps, la plus grande intensité est mesurée au printemps (5,9 UTN le 6 avril 1998), alors que l'automne monte jusqu'à un maximum de 1,32 UTN (le 8 décembre 2003).

6.1.4 Analyse statistique de TURF_ATW

Les statistiques descriptives sont données au Tableau 6-1 pour toutes les données, et pour chaque saison.

Tableau 6-1 : Statistiques descriptives de TURF_ATW, pour l'année et par saison

	TURF_ATW (en UTN)									
	N	Moyenne	Médiane	Minimum	Maximum	P25	P75	P90	P95	Écart type
Toutes les données	3412	0,23	0,17	0,05	5,90	0,12	0,26	0,40	0,52	0,24
Printemps	570	0,41	0,26	0,08	5,90	0,20	0,42	0,80	1,12	0,49
Été	1050	0,14	0,12	0,05	0,65	0,11	0,14	0,20	0,28	0,06
Automne	1792	0,23	0,20	0,05	1,32	0,14	0,27	0,39	0,49	0,14

Sur les 3412 exemples disponibles, toutes saisons confondues, la qualité générale de l'eau est excellente : 95 % des valeurs sont inférieures ou égales à 0,52 UTN. De nouveau, le printemps représente la saison où l'eau est la plus difficile à traiter. 10% des 570 valeurs qui le composent sont supérieures à 1,12 UTN. Le maximum de turbidité est atteint le 6 avril 1998 avec 5,90 UTN, soit deux jours après le maximum de turbidité mesuré à l'eau filtrée de Des Baillets. Le décalage temporel de deux journées entre les deux usines apparaît de nouveau ici. Une illustration graphique du tableau ci-dessus est donnée à la Figure 6-2.

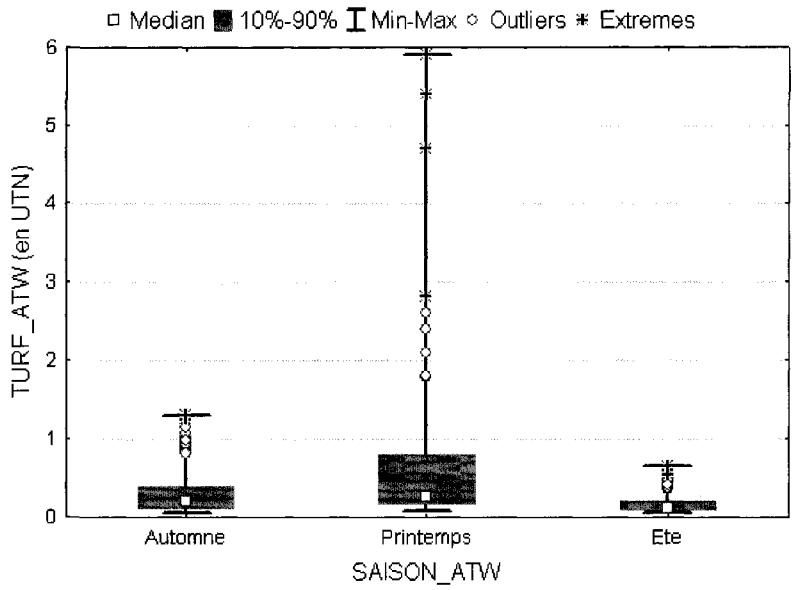

Figure 6-2 : Variations saisonnières de la turbidité à l'eau filtrée de l'usine Atwater - TURF_ATW

6.1.5 Partitionnement des exemples

Afin de pouvoir joindre la prédiction de l'eau filtrée à celle de l'eau brute dans une même interface prédictive, le partitionnement des exemples utilisé pour la prédiction à l'eau brute d'Atwater sera de nouveau employé ici. Il comprend deux répartitions (ATW 1 et 2) issues d'un échantillonnage aléatoire à proportion fixée parmi des classes de turbidité.

6.1.6 Choix des entrées

Analyse des corrélations

Tableau 6-2 : Corrélations croisées de TURF_ATW avec les variables de qualité de l'eau

	TURF_ATW			
	Toute l'année	Automne	Été	Printemps
TEMPEB_DB-1	-0,35	-0,42	-0,15	-0,27
COUL_DB-2	0,60	0,64	0,22	0,69
TURB_DB-2	0,44	0,32	-0,08	0,84
COUL_ATW-1	0,71	0,62	0,15	0,87
TURB_ATW	0,72	0,61	0,12	0,89
TURB_ATW-1	0,56	0,69	0,23	0,63
TURF_ATW-1	0,93	0,88	0,70	0,94

Quelle que soit la saison, il y a une forte corrélation entre TURF_DB et sa valeur la veille (r compris entre 0,70 et 0,94).

L'automne, les variables les plus corrélées sont par ordre décroissant : TURF_ATW-1, TURB-ATW-1, COUL_DB-2, et COUL_ATW-1. La couleur à Des Baillets deux jours avant s'avère importante, elle peut être un indicateur de la qualité de l'eau le jour même à Atwater.

Pour l'été, les variables TURF_ATW-1, TURB_ATW-1 et COUL_DB-2 sont retenues.

Le printemps est la saison montrant les plus hauts coefficients de corrélation. COUL_ATW-1 est retenu à la place de TURB_ATW-1 car il montre une meilleure corrélation. TURB_DB-2 et COUL_DB-2 seront également des candidats potentiels car ils peuvent aussi refléter la qualité de l'eau le jour même à Atwater.

Le modèle linéaire annuel se basera quant à lui sur les variables TURF_ATW-1, COUL_DB-2, COUL_ATW-1, et IDX_SAISON_ATW.

6.1.7 Tableau récapitulatif des modèles retenus

Les entrées des modèles et l'architecture de ces derniers ont été déterminées par l'algorithme IPS de Statistica. Les paramètres internes utilisés sont les mêmes que pour le modèle régressif du Chapitre 3. Les modèles finaux retenus sont décrits dans le Tableau 6-3.

Il ressort que les modèles pour l'automne et le printemps ont conservé les quatre variables de qualité initialement candidates (TURF_ATW-1, TURB_ATW-1, TEMPEB_DB-1, et COUL_DB-2 ou TURB_DB-2), alors que pour l'été deux variables (TURF_ATW-1 et TURB_ATW-1) sont retenues. De plus, l'automne et le printemps incluent chacune une variable représentative de la dégradation de la turbidité à l'eau brute, LSF_VITM-3 et RIV_CHAT-5 respectivement.

Tableau 6-3 : Résumé des modèles prédictif retenus pour TURF_ATW

	Annuel	Saisonnier		
		Automne	Printemps	Été
Type	Linéaire	PMC	PMC	PMC
	4:1	5:7:1	5:6:1	2:8:1
Entrées	TURF_ATW-1	TURF_ATW-1	TURF_ATW-1	TURF_ATW-1
	TURB_ATW-1	TURB_ATW-1	TURB_ATW-1	TURB_ATW-1
	COUL_DB-2	TEMPEB_DB-1	TEMPEB_DB-1	
	SAISON_ATW	COUL_DB-2	TURB_DB-2	
		LSF_VITM-3	RIV_CHAT-5	

6.2 Résultats

Sur le Tableau 6-4, sont affichés les résultats des modèles linéaires annuels et PMC saisonniers pour les deux répartitions (ATW 1 et 2). Les trois critères de performance observés sont toujours : r, EAM, et EQM (si la 'saison' considérée présente beaucoup de valeurs supérieures ou égales à 0,8 UTN). Les performances du modèle linéaire sont comparées à celles des modèles saisonniers (en italique sur la deuxième ligne), le plus performant des deux étant inscrit en gras.

Tableau 6-4 : Comparaison des performances des modèles annuels et saisonniers pour la prévision de TURF_DB

Répartition ATW 1	Toutes les données			Automne		Printemps			Été	
	r	EQM	EAM	r	EAM	r	EQM	EAM	r	EAM
Test	0,927	0,059	0,033	0,890	0,032	0,927	**0,059**	**0,033**	0,706	0,029
				0,937	***0,025***	***0,942***	*0,075*	*0,043*	***0,829***	***0,020***
Toutes les données	0,938	0,078	0,038	0,898	0,033	0,938	**0,078**	**0,038**	0,672	0,030
				0,914	***0,030***	***0,978***	*0,092*	*0,053*	***0,781***	***0,023***
TURF_DB obs >= 0,8 UTN	0,894	0,39	0,25	**0,746**	0,22	0,896	0,41	0,26	Aucun exemple >= 0,8UTN	
				0,698	***0,21***	***0,961***	***0,246***	***0,18***		
	Linéaire 4:1			*PMC 5:7:1*		*PMC 5:6:1*			*PMC 2:8:1*	

Répartition ATW 2	Toutes les données			Automne		Printemps			Été	
	r	EQM	EAM	r	EAM	r	EQM	EAM	r	EAM
Test	0,944	0,054	0,034	**0,920**	**0,031**	0,944	**0,054**	**0,034**	0,509	0,030
				0,914	*0,033*	***0,974***	*0,071*	*0,061*	***0,863***	***0,024***
Toutes les données	0,938	0,078	0,038	0,898	0,033	0,938	**0,078**	**0,038**	0,669	0,030
				0,900	*0,033*	***0,959***	*0,13*	*0,075*	***0,792***	***0,023***
TURF_DB obs >= 0,8 UTN	0,894	0,39	0,25	**0,748**	**0,23**	0,896	0,41	**0,26**	Aucun exemple >= 0,8UTN	
				0,694	*0,27*	***0,924***	***0,37***	*0,28*		
	Linéaire 4:1			*PMC 5:7:1*		*PMC 5:6:1*			*PMC 2:8:1*	

Il résulte que le modèle saisonnier d'été apporte une amélioration notable du coefficient de corrélation, ceci sur tous les ensembles de données. Par exemple, pour la répartition ATW1, r passe de 0,71 à 0,83 sur les données test. Aucun dépassement supérieur à 0,8 UTN n'est observé, c'est pourquoi la rubrique est vide.

En ce qui concerne le printemps, sur toutes les données et l'ensemble test, les critères EQM et EAM semblent privilégier le modèle linéaire annuel. Cependant, la différence est faible : de l'ordre de quelques centièmes d'UTN. Le coefficient de corrélation (r) se trouve amélioré systématiquement de +0,02 à +0,03 unités. L'étude des évènements observés supérieurs à 0,8 UTN favorise plus clairement le modèle neuronal avec une baisse cumulée de l'EQM jumelée à une augmentation de r.

En revanche, pour l'automne, même si le modèle neuronal donne des performances quasi équivalentes ou meilleures que le modèle linéaire, ses performances se dégradent sur la prédiction des évènements supérieurs à 0,8 UTN. En effet, la

corrélation perd systématiquement 0,05 unités. Cette baisse de performance s'explique par le faible nombre de données supérieures à 0,8 UTN pour TURF_ATW : l'automne, pris séparément, contient trop peu d'exemples de 'haute turbidité' pour pouvoir capturer efficacement la loi physique sous-jacente. Le modèle annuel, bien que linéaire, dispose quant à lui de tous les exemples 'hauts'.

Sur la Figure 6-3 et la Figure 6-4, sont tracées les prédictions données par les modèles saisonniers neuronaux en fonction des valeurs observées.

Pour le printemps (Figure 6-3), une bonne concordance est obtenue sur toute la gamme de turbidités observées : la régression linéaire des points expérimentaux (en pointillés) est quasiment confondue avec la droite d'équation y=x (en trait d'axe). Un seul point est prédit inférieur à 0,8 UTN sur les 36 points observés supérieurs à 1 UTN, soit environ 3% de mauvaise classification des évènements 'hauts'.

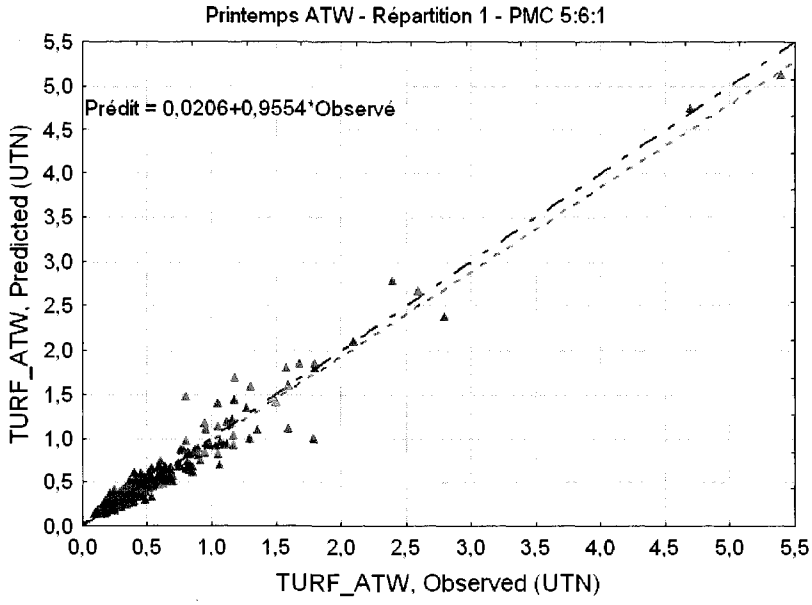

Figure 6-3 : TURF_ATW prédite en fonction d'observée-printemps-répartitionATW1-toutes les données

Concernant les résultats de l'automne, la Figure 6-4 montre des prédictions satisfaisantes de 0 à 0,5 UTN. En revanche, à partir des données observées supérieures à 0,5 UTN, de nombreux cas se retrouvent sous-estimés par le modèle neuronal : à 1 UTN observé, la droite de régression linéaire (en pointillés) est environ 0,2 UTN en deçà de la courbe y=x (trait d'axe). Cependant, malgré cette sous-estimation, sur les 5 points observés supérieurs à 1 UTN, aucun n'est prédit inférieur à 0,8 UTN. En termes d'opération de la station, ceci présente des performances suffisantes qui tendraient à valider le modèle neuronal de l'automne.

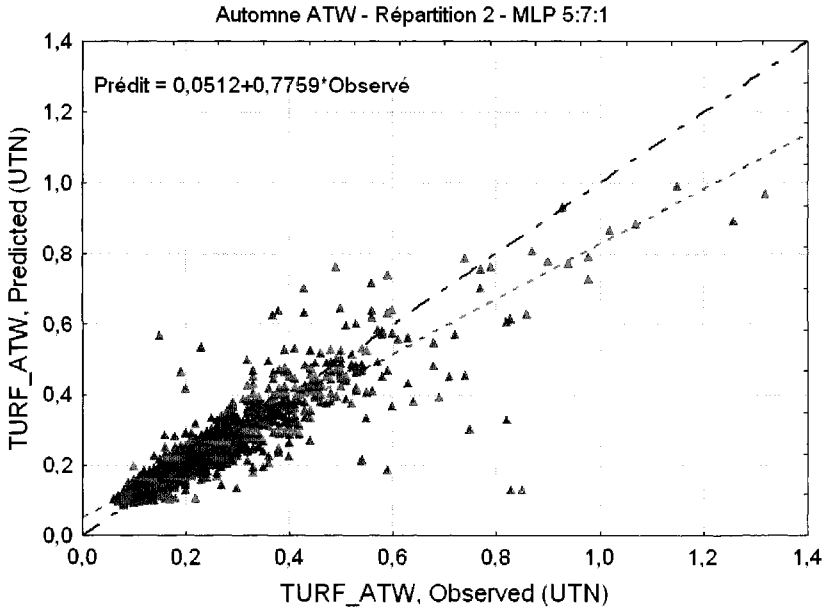

Figure 6-4 : TURF_ATW prédite en fonction d'observée-automne-répartitionATW2-toutes les données

6.3 Discussion

6.3.1 Améliorations des prédictions à l'automne et au printemps

À la vue des résultats présentés ci-dessus, le printemps est suffisamment bien prédit par le modèle neuronal (environ 3% de mauvaise classification).

Le modèle pour l'automne, par contre, pourrait être amélioré particulièrement pour les prédictions d'évènements supérieurs à 0,8 UTN par l'implantation d'un modèle de classification. Il conviendrait de tester les deux pistes suivantes pour augmenter le nombre d'exemples « hauts » disponibles et améliorer les prédictions obtenues:

- Inclure tous les évènements supérieurs à 0,8 UTN dans un modèle de prédiction annuel de type PMC.

- Abaisser le seuil de turbidité, au prix d'un peu plus de chevauchement des classes.

La première option va à l'encontre des résultats observés sur les facteurs explicatifs saisonniers. La deuxième option serait donc à privilégier pour viser de meilleurs résultats.

La remarque de la Section 5.3.2 s'applique aussi ici : l'inclusion des charges superficielles sur les filtres ou bien le débit d'eau journalier produit par l'usine pourrait améliorer les prédictions. Ces valeurs affectent directement la qualité de l'eau en sortie des filtres. Compte tenu des données disponibles pour le projet, elles n'ont pas été incluses pour ne pas renoncer à cinq années d'exemples (notamment l'année 1998).

Tenir compte des valeurs prédites pour la turbidité à l'eau brute (TURB_ATW) ne devrait pas augmenter significativement les performances des modèles saisonniers. En effet, l'usine Atwater dispose des données de qualité d'eau en amont à Des Baillets. La qualité de l'eau brute à Atwater est fortement corrélée avec celle de Des Baillets deux jours avant. Ainsi les variables de qualité COUL_DB-2 et TURB_DB-2 ont déjà été incluses dans l'élaboration des modèles pour représenter la qualité de l'eau brute TURB_ATW.

CONCLUSION ET PERSPECTIVES

Conclusion

Ce projet avait pour but de remplir quatre objectifs consistant en l'élaboration de quatre modèles prédictifs des pointes de turbidité à l'eau brute et à l'eau filtrée pour les usines Des Baillets et Atwater. Les codes de ces modèles étalonnés ont été stockés en format « Statistica Visual Basic® » en vue d'une possible implantation au sein d'une base de données de type Access®.

Une méthodologie de construction d'un modèle par réseau de neurones a été définie pour répondre aux besoins du problème. Cette méthodologie peut être transférée à d'autres situations. De plus, des techniques novatrices telles que le prétraitement des données par fonction de répartition ont été essayés ici et ont fait leurs preuves pour certains exemples particuliers.

La continuité des travaux de Tremblay (2004) permit de corroborer certains résultats quant à la détermination des variables influentes pour la prédiction de la turbidité à l'eau brute de Des Baillets. En effet, le recensement des 213 évènements turbides (≥3.1 UTN) et l'occurrence des huit causes expliquant potentiellement ces hausses ont fait ressortir que :

- À l'automne, les pointes de turbidité sont principalement sous influence directe des tempêtes de vent; tempêtes responsables de la remise en suspension des sédiments des lacs Saint-François et Saint-Louis. Durant les mois de janvier – février, la présence d'un couvert de glace offre une protection naturelle à ces intempéries. Or, lors d'épisodes de chaleur hivernale accompagnée de pluie, la fragilité ou bris de cette protection naturelle met à nu le lit du lac qui est particulièrement vulnérable du fait de l'absence d'herbier aquatique. Ceci expliquerait les évènements isolés de la fin de l'automne et du début du printemps.

- Au printemps, une des causes majeures est la fonte des neiges impliquant le déversement des eaux au barrage de Carillon et l'augmentation du rapport de mélange des Outaouais par rapport au fleuve Saint-Laurent. De plus, la fonte des neiges pourrait être indirectement responsable du ruissellement et de la hausse des tributaires secondaires. Le rapport de mélange utilisé ici ne tenait compte que de la proportion de la rivière Outaouais passant par le chenal de Vaudreuil. Ce rapport s'est avéré être un indicateur important pour la prédiction. Ceci renforce l'hypothèse que la prise d'eau de Montréal serait sous influence limitée des eaux de mélange provenant de Sainte-Anne de Bellevue, et qu'elle dépendrait grandement du mélange dans la gire de l'île Perrot. Les eaux de la gire suivraient ensuite le chenal de navigation ou les vitesses d'écoulement sont plus élevées, rendant la sédimentation des particules plus difficiles.
- Le renversement au lac des Deux Montagnes contribue à la dégradation de la qualité moyenne de l'eau au printemps et à l'automne sans pour autant expliquer les fortes hausses de turbidité.
- Les précipitations n'influencent qu'épisodiquement la turbidité à l'eau brute. Elles semblent moins importantes que le facteur vent.

En ce qui concerne les modèles neuronaux développés et la méthodologie abordée :

- La mise en évidence de comportements saisonniers pouvant obéir à des lois physiques différentes, et l'utilisation de modèles spécifiques saisonniers permet d'améliorer la performance du réseau créé.
- De même le découpage horizontal en modèles de classification pour divers seuils de turbidité permet de faire ressortir les causes explicatives propres à chaque intensité d'évènement turbide. Par exemple, les causes renversement et tempête de vent n'ont pas la même intensité à l'automne. Le premier dégrade

la valeur moyenne de l'eau sur plusieurs jours alors que le second peut causer des pointes de forte amplitude durant une courte durée.

- Le tri et l'analyse de la base de donnée permet d'enlever les informations non pertinentes, de gagner en connaissance sur le phénomène à modéliser, et permet ultérieurement de définir les causes des évènements mal classés par un modèle.
- Le partitionnement des exemples gagne à être effectué en préservant la représentativité des données dans chaque ensemble, qui doivent être issues de la même population statistique. Ainsi, l'utilisation de dates fixes est à proscrire.
- Le choix des entrées combine différentes approches. Soit dans l'ordre : connaissances préalables du modélisateur sur le sujet, inspection visuelle du phénomène, analyses statistiques linéaires (corrélations et analyse discriminante), et au besoin l'étude des performances de modèles neuronaux de type GRNN calibrés avec différentes combinaisons des entrées potentielles.
- Le critère de performance utilisé doit être adopté aux besoins du modèle. Il est essentiel de bien définir les objectifs auxquels il doit répondre, et au besoin de se créer un ou plusieurs critères de performance représentatifs de ces besoins.

Concernant l'étude sur l'usine Atwater, il ressort que :

- L'eau brute à l'usine exhibe une forte corrélation avec la qualité de l'eau brute à Des Baillets deux jours avant, et ce quelle que soit la saison.
- L'eau brute reste sous faible influence des précipitations. À moins d'avoir des pluies exceptionnelles, il semble ne pas y avoir d'impact notable du ruissellement urbain sur la qualité de l'eau dans le canal.
- Les quelques 8 km de canal agissent en tant que décanteur, ainsi la turbidité à l'eau brute observée à Atwater et beaucoup moins forte qu'à Des Baillets.
- La saison printanière reste une saison critique. Là où l'automne présente des pointes de très forte amplitude à l'automne à Des Baillets, ces pointes se

retrouvent estompés rendus à Atwater. Les particules responsables de la turbidité doivent être assez grosses pour décanter dans le canal.

- En revanche, le printemps est la saison critique pour l'eau brute à Atwater avec des pointes pouvant atteindre 15 UTN. Le canal semble n'avoir qu'un faible impact sur les pointes du printemps.

En termes de traitement et de turbidité en sortie des filtres :

- Le printemps est la saison critique où les deux usines peuvent expérimenter des difficultés de filtration. Ceci doit aussi être relié à une nature de particules différentes au printemps et à l'automne.
- La turbidité en sortie des filtres variant lentement, elle est assez bien prédite par sa valeur de la veille. Il suffit d'inclure au modèle des variables de qualité de l'eau brute et, au besoin, des variables explicatives des pointes saisonniers pour obtenir une bonne prédiction de la turbidité à l'eau filtrée.

Il est intéressant de noter que le printemps est la période critique pour le traitement et l'eau brute à Atwater, sans doute à cause d'une nature de particules différente. Lors du déversement des eaux du barrage de Carillon, le rapport de mélange des Outaouais passe de l'ordre de 3 à 6 % à 20 % environ. Les sols en amont de l'Outaouais sont fortement occupés par des terres agricoles et autres élevages. L'eau est sans doute plus chargée en matière organique naturelle et autres substances colloïdales de faible taille, ou même des substances dissoutes dans l'eau. D'où la difficulté à traiter l'eau avec des filtres à sable non assistés chimiquement et la dégradation de la couleur observée chaque printemps par les opérateurs de l'usine de Pointe Claire.

Dans les années à venir, une plus grande contribution de la rivière des Outaouais par rapport au fleuve Saint-Laurent est annoncée par les experts en changement climatique. Ce rapport de mélange accru pourrait être responsable de la dégradation de la qualité moyenne à l'eau brute pour la ville de Montréal.

Perspectives

Les codes des modèles développés ont été stockés en langage Visual Basic en vue de leur implantation en station. Ceci serait réalisable au moyen d'une base de données de type Access. Dans un premier temps, à chaque jour, les données pourraient être récupérées par des systèmes d'acquisition automatique (capteur télémétrique, données disponibles sur internet, etc.). Ensuite, une interface graphique fournirait toutes les informations utiles à l'opérateur pour prendre une décision. À savoir :

- Une courbe avec la turbidité des vingt derniers jours observés accompagnée de leur prédiction. Ceci afin de situer la tendance de la courbe de turbidité.
- Les résultats des modèles de classification en cascade et de régression pour la saison dont l'indice de pertinence est le plus élevé.
- En cas de chevauchement des saisons, ces résultats seront accompagnés de ceux de la saison dont l'indice de pertinence est inférieur.
- Pour avoir un jugement conservateur, les indices de renversement et de fonte des neiges des 10 derniers jours devraient être affichés en tant qu'alarme. Leur activation pousserait l'opérateur à choisir la solution la plus conservatrice si deux prédictions sont disponibles.

La construction d'un tel système permettrait d'alimenter en continu une base de données et de re-calibrer les modèles au bout de quelques années. Ceci permettrait de prendre en compte une éventuelle dérive des valeurs mesurées, et sans doute d'améliorer la performance des modèles retenus en les ré-entraînant avec plus d'exemples.

BIBLIOGRAPHIE

BAXTER, C.W., SHARIFF, R., STANLEY, S.J., SMITH, D.W., ZHANG, Q., SAUMER, E.D. 2002. Model-based advanced process control of coagulation. *Water Science and Technology.* 45 : 4-5. 9-17.

BAXTER, C.W., ZHANG, Q., STANLEY, S.J., SHARIFF, R., TUPAS, R.-R.T., STARK, H.L. 2001. Drinking water quality and treatment: the use of artifical neural networks. *Canadian Journal of Civil Engineering.* 28 : Suppl. 1. 26-35.

BEAUDEAU, P., LEBOULANGER, T., LACROIX, M., HANNETON, S., WANG, H.Q. 2001. Forecasting of turbid floods in a coastal, chalk karstic drain using an artificial neural network. *Ground Water.* 39 : 1. 109-118.

BIRIKUNDAVYI, S., LABIB, R., TRUNG, H.T., ROUSSELLE, J. 2002. Performance of neural networks in daily streamflow forecasting. *Journal of Hydrologic Engineering.* 7 : 5. 392-398.

BISHOP, C.M., 1995. *Neural networks for pattern recognition (First Edition).* Oxford (Angleterre), Clarendon Press, Oxford University Press. 482 p.

BOWDEN, G.J., DANDY, G.C., MAIER, H.R. 2005. Input determination for neural network models in water resources applications. Part 1 - Background and methodology. *Journal of Hydrology.* 301 : 1-4. 75-92.

BRION, G.M., LINGIREDDY, S. 2003. Artificial neural network modelling: a summary of successful applications relative to microbial water quality. *Water Science and Technology.* 47 : 3. 235-240.

CHAMPOUX, L., SLOTERDIJK, H. H. 1988. *Étude de la qualité des sédiments du lac Saint-Louis 1984-1985 (Rapport technique no. 1).* Environnement Canada, Conservation et Protection, Région du Québec. Québec, Canada.

CIGIZOGLU, H.K., 2002a. Suspended sediment estimation and forecasting using artificial neural networks. *Turkish Journal of Engineering and Environmental Sciences.* 26 : 1. 15-25.

CIGIZOGLU, H.K., 2002b. Suspended sediment estimation for rivers using artificial neural networks and sediment rating curves. *Turkish Journal of Engineering and Environmental Sciences.* 26 : 1. 27-36.

CIGIZOGLU, H.K., 2004. Estimation and forecasting of daily suspended sediment data by multi-layer perceptrons. *Advances in Water Resources.* 27 : 2. 185-195.

CIGIZOGLU, H.K., 2005. Application of generalized regression neural networks to intermittent flow forecasting and estimation. *Journal of Hydrologic Engineering.* 10 : 4. 336-341.

CIGIZOGLU, H.K., ALP, M. 2006. Generalized regression neural network in modelling river sediment yield. *Advances in Engineering Software.* 37 : 2. 63-68.

CIGIZOGLU, H.K., KISI, O. 2006. Methods to improve the neural network performance in suspended sediment estimation. *Journal of Hydrology.* 317 : 3-4. 221-238.

CORANI, G., GUARISO, G. 2005. An application of pruning in the design of neural networks for real time flood forecasting. *Neural Computing and Applications.* 14 : 1. 66-77.

COUILLARD, D., 1987. Qualité des sédiments en suspension et de fond du système Saint-Laurent (Canada). *Journal des Sciences Hydrologiques.* 32 : 4. 445-464.

COULIBALY, P., ANCTIL, F., BOBÉE, B. 1999. Prévision hydrologique par réseaux de neurones artificiels: état de l'art. *Canadian Journal of Civil Engineering.* 26 : 3. 293-304.

DREYFUS, G., SAMUELIDES, M., MARTINEZ, J.-M., GORDON, M.B., BADRAN, F., THIRIA, S., HÉRAULT, L. 2004. *Réseaux de neurones: méthodologies et applications (Deuxième Édition)*. Eyrolles. Paris, France. 417 p.

FORTIN, G.R., AUCLAIR, M.-J., LETIENNE-PRÉVOST, M., PORTIN, P., SÉGUIN, D. 1994. *Synthèse des connaissances sur les aspects physiques et chimiques de l'eau et des sédiments du lac Saint-Louis (Rapport technique). Zone d'intervention prioritaire*. Environnement Canada. Région du Québec, Conservation de l'Environnement, Centre Saint-Laurent. Québec, Canada. 177 p.

FRENETTE, R., FRENETTE, M. (1992). *Modélisation des bilans sédimentaires du Saint-Laurent tronçon aval : Montréal-Montmagny* (modèle Bi-Lavsed). Congrès annuel de la société canadienne de génie civil, Mai 27-29, 1992, Québec.

GAGNON, C., GRANDJEAN, B.P.A., THIBAULT, J. 1997. Modelling of coagulant dosage in a water treatment plant. *Artificial Intelligence in Engineering.* 11 : 4. 401-404.

GOVINDARAJU, R.S., 2000a. Artificial neural networks in hydrology. I: Preliminary concepts. *Journal of Hydrologic Engineering.* 5 : 2. 115-123.

GOVINDARAJU, R.S., 2000b. Artificial neural networks in hydrology. II: Hydrologic applications. *Journal of Hydrologic Engineering.* 5 : 2. 124-137.

HAYKIN, S., 1999. *Neural networks: a comprehensive foundation (Second Edition).* Pearson and Prentice Hall. Upper Saddle River, New Jersey, USA. 842 p.

HERNANDEZ, H., LE LANN, M.-V. 2006. *Development of a neural sensor for on-line prediction of coagulant dosage in a potable water treatment plant in the way of its diagnosis*. Springer Verlag, Heidelberg, Germany. Ribeirao Preto, Brazil.

HORNIK, K., STINCHCOMBE, M., WHITE, H. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks.* 2 : 5 . 359-366.

KINGSTON, G.B., MAIER, H.R., LAMBERT, M.F. 2005. Calibration and validation of neural networks to ensure physically plausible hydrological modeling. *Journal of Hydrology.* 314 : 1-4. 158-176.

LORRAIN, S., PELLETIER, M., FORTIN, G. 1999. *Synthčse des connaissances sur les aspects physiques et chimiques de l'eau et des sédiments du lac Saint-Louis: addenda. Nord du lac Saint-Louis: zone d'intervention prioritaire 5 Montréal.* Centre Saint-Laurent, Conservation de l'Environnement, Environnement Canada, Région du Québec. Québec, Canada. 58 p.

MAIER, H.R., DANDY, G.C. 1997. Determining inputs for neural network models of multivariate time series. *Microcomputers in Civil Engineering.* 12 : 5. 353-368.

MAIER, H.R., DANDY, G.C. 1998a. The effect of internal parameters and geometry on the performance of back-propagation neural networks: an empirical study. *Environmental Modelling and Software.* 13 : 2. 193-209.

MAIER, H.R., DANDY, G.C. 1998b. Understanding the behaviour and optimizing the performance of back-propagation neural networks: an empirical study. *Environmental Modelling and Software.* 13 : 2. 179-191.

MAIER, H.R., DANDY, G.C. 1999. Empirical comparison of various methods for training feed-forward neural networks for salinity forecasting. *Water Resources Research.* 35 : 8. 2591-2596.

MAIER, H.R., DANDY, G.C. 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling and Software.* 15 : 1. 101-124.

MAIER, H.R., DANDY, G.C. 2001. Neural network based modelling of environmental variables: a systematic approach. *Mathematical and Computer Modelling.* 33 : 6-7. 669-682.

MAIER, H.R., MORGAN, N., CHOW, C.W.K. 2004. Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters. *Environmental Modelling and Software.* 19 : 5. 485-494.

MARTINEC, J., RANGO, A. 1989. Merits of statistical criteria for the performance of hydrological models. *Water Resources Bulletin.* 25 : 2. 421-432.

MOISE, A., SALAMON, R., COMMENGES, D., CLÉMENT, B. 1986. L'utilisaton des courbes ROC. *Revue Épidémiologique et Santé Publique.* 34 : 3. 209-217.

NEELAKANTAN, T.R., BRION, G.M., LINGIREDDY, S. 2001. Neural network modelling of Cryptosporidium and Giardia concentrations in the Delaware River, USA. *Water Science and Technology.* 43 : 12. 125-132.

NIST/SEMATECH, 2006. *e-Handbook of Statistical Methods.* [En ligne]. Tiré de : http://www.itl.nist.gov/div898/handbook/index.htm (Page consultée le 11/7/2007).

NOUR, M.H., SMITH, D.W., EL-DIN, M.G., PREPAS, E.E. 2006a. The application of artificial neural networks to flow and phosphorus dynamics in small streams on the Boreal Plain, with emphasis on the role of wetlands. *Ecological Modelling.* 191 : 1. 19-32.

NOUR, M.H., SMITH, D.W., GAMAL EL-DIN, M., PREPAS, E.E. 2006b. Neural networks modelling of streamflow, phosphorus, and suspended solids: application to the Canadian Boreal forest. *Water Science and Technology.* 53 : 10. 91-99.

ÖZESMI, S.L., TAN, C.O., ÖZESMI, U. 2006. Methodological issues in building, training, and testing artificial neural networks in ecological applications. *Ecological Modelling.* 195 : 1-2. 83-93.

RECKNAGEL, F., 2001. Applications of machine learning to ecological modelling. *Ecological Modelling.* 146 : 1-3. 303-310.

RONDEAU, B., COSSA, D., GAGNON, P., BILODEAU, L. 2000. Budget and sources of suspended sediment transported in the St. Lawrence River, Canada. *Hydrological Processes.* 14 : 1. 21-36.

SHAHIN, M.A., MAIER, H.R., JAKSA, M.B. 2004. Data division for developing neural networks applied to geotechnical engineering. *Journal of Computing in Civil Engineering.* 18 : 2. 105-114.

SNC-PROCÉAN (1992). *Caractérisation physico-chimique des sédiments du lac Saint-Louis*. Environnement Canada, Conservation et Protection, région du Québec, Centre Saint-Laurent, rapport d'étude-pilote.

STATSOFT, 2006. *Electronic statistics textbook.* [En ligne]. Tiré de : http://www.statsoft.com/textbook/stathome.html (Page consultée le 11/7/2007).

SUDHEER, K.P., JAIN, A., SRINIVASULU, S. 2004. Discussion of "performance of neural networks in daily streamflow forecasting". *Journal of Hydrologic Engineering.* 9 : 6. 553-555.

THE MATHWORKS, 2007. *Matlab - Neural network toolbox help file.* Version 5.1. [Logiciel]. Natik, Massachusetts, USA.

TREMBLAY, G., 2004. *Prévision des augmentations de turbidité à l'eau brute de la Ville de Montréal par des réseaux de neurones artificiels (Maîtrise).* École Polytechnique - Génie Civil, Géologique et des Mines. Québec, Canada. 248 p.

VALENTIN, N., DENOEUX, T., FOTOOHI, F. 1999. *An hybrid neural network based system for optimization of coagulant dosing in a water treatment plant.* International Joint Conference on Neural Networks. Washington, DC, USA.

YU, R.-F., KANG, S.-F., LIAW, S.-L., CHEN, M.-C. 2000. Application of artificial neural network to control the coagulant dosing in water treatment plant. *Water Science and Technology.* 42 : 3-4. 403-408.

ZHANG, Q., STANLEY, S.J. 1999. Real-time water treatment process control with artificial neural networks. *Journal of Environmental Engineering.* 125 : 2. 153-160.

A - ANNEXES

Annexe A Définition des variables d'index

Afin de représenter les évènements de fonte de neiges et de renversement, deux variables d'index appelées respectivement IDX_FONT et IDX_RENV furent construites. Sont aussi présentées les informations de construction relative à la contribution des Outaouais dans le fleuve Saint-Laurent, et un indicateur de la pluie maximale autour de l'île de Montréal.

Index de fonte des neiges IDX_FONT

L'objectif de cette section est de refléter la fonte printanière, ou tout simplement un évènement de fragilisation du couvert de glace à cause du réchauffement couplé à des précipitations. Ainsi, un index indicateur de la fonte des neiges fut créé, index basé sur les débits hydrologiques, sur le débit au barrage de Carillon et sur la température de l'air moyenne entre les stations météorologiques situées à Dorval, Sainte-Anne de Bellevue, et Lac Saint-François.

Cet index de fonte ne peut prendre que 3 valeurs discrètes [0; 1/2; 1]. La valeur 1 symbolisant la fonte des neiges et le déversement important d'eau au barrage de Carillon sur la rivière des Outaouais, cette valeur traduit souvent le bris total du couvert de glace. La valeur ½ étant un état intermédiaire où les tributaires secondaires augmentent au-delà de valeurs seuils fixées à cause de précipitations abondantes et du réchauffement de la température de l'air, cette valeur traduit souvent la fragilité du couvert de glace. La valeur 0 étant active le reste du temps.

À partir des valeurs seuils d'activation de causes, valeurs suggérées par Tremblay (2004), et en fonction des valeurs moyennes des variables au printemps, des valeurs seuils furent déterminées de telle sorte que celles-ci soient proches de la médiane des

données (arrondies à la valeur inférieure) : à savoir 5, 9, 50, et 2400 m^3/d pour les débits des rivières Beaudette, des Raisins, de Châteauguay, et du barrage de Carillon respectivement.

Pour traduire l'effet d'inertie dû au déversement des masses d'eau stockées par le barrage de Carillon, un « terme de mémoire » fut ajouté. Lorsque l'index de fonte passe à 1, tant que le débit au barrage est croissant, l'index reste à 1 et la valeur maximale atteinte par ce débit est stockée. Si OUT_CARI (voir les notations des variables au Tableau 3-1) décroît en dessous de 95% de son maximum, l'index perd la valeur 1, il peut reprendre ses valeurs normales (soit 0 ou ½). La valeur 95% a été déterminée par essais et erreurs en observant graphiquement les pointes de turbidité et IDX_FONT en fonction du temps. Les zones du graphique où IDX_FONT était actif (½ ou 1) devaient englober les pointes de turbidité.

On exprime les conditions de passage à 0, ½ ou 1 de la manière suivante :

- conditions de passage à ½ : deux des trois tributaires secondaires (Raisin, Beaudette ou Châteauguay) doivent dépasser leur seuil ET la température moyenne de l'air doit être positive les deux jours précédents
- conditions de passage à 1 :
    - trois tributaires dépassent leur seuil ET dépassement du seuil à Carillon ET température moyenne de l'air positive les 6 derniers jours
    - OU si IDX_FONT_lag1 est à 1, et que le débit à Carillon est au dessus de 95% du débit max pendant le pic étudié alors IDX_FONT reste à 1 SINON il prend la valeur qu'il devrait prendre normalement
- 0 sinon

Notons que la température de l'air est assez stable dans les diverses stations à un jour donné; ces données sont regroupées en faisant la moyenne quotidienne de la température de l'air aux stations de Dorval, Sainte-Anne de Bellevue et lac Saint-François.

Index de renversement IDX_RENV

Le renversement est un phénomène mettant en mouvement les masses d'eau d'un lac. Sous réserve que le lac soit stratifié, lorsque la température de l'eau atteint 4°C (température du maximum de densité de l'eau), il se produit une circulation de l'eau des couches supérieures vers les couches inférieures. La stratification peut avoir lieu si : le lac est suffisamment profond, s'il ne fonctionne pas comme un réacteur complètement mélangé (dû à l'arrivée de nombreux affluents et à de fortes turbulences). Au fond du lac, s'accumule une couche d'eau froide appelée hypolimnion. Cette couche est en contact direct avec les sédiments accumulés au fil du temps (métaux, argiles et sables, nutriments re-largués par la biomasse sédimentée et en décomposition). Lors du brassage ces sédiments peuvent être remis en suspension et être une cause d'évènements turbides.

Ce phénomène de renversement peut arriver deux fois par an : à l'automne et au printemps. À la fin de l'automne, les eaux de surface se refroidissent, deviennent plus denses, descendent et remplacent la couche inférieure. Durant l'hiver la température au sein du lac est assez homogène (inférieure à 4°C) jusqu'au réchauffement printanier. La couche de surface se réchauffant après la fonte du couvert de glace, la température de l'eau de surface remonte en passant par 4°C, il se produit un deuxième brassage des couches d'eau.

Concernant le système étudié, il y a potentiellement trois lacs où pourrait se produire le renversement : le lac Saint-François, le lac Saint-Louis, et le lac des Deux Montagnes. Seul ce dernier est sujet à la stratification, donc au renversement.

L'estimation du renversement au lac des Deux Montagnes ce fait en utilisant un indicateur double : la température quotidienne moyenne de l'eau brute à Hawkesbury, et celle de l'eau filtrée à Oka. Le premier donne une idée de la température plus en amont sur la rivière des Outaouais, puis la deuxième station, renseigne sur la situation juste devant le lac. En effet, le lac constituant une forte masse d'eau, la température de celle-ci ne risque pas de varier aussi vite que l'eau des Outaouais. Ainsi, prendre seulement un indicateur situé loin en amont du lac risquerait d'anticiper le renversement. Le choix de la ville d'Hawkesbury s'est fait à cause de la disponibilité des données. L'inclusion de la température à Oka permet de confirmer que les masses d'eau intermédiaires entre Hawkesbury et le lac n'ont pas joué un effet tampon.

Pour représenter le flou existant sur le vrai moment du renversement par rapport à celui suggéré par la température de l'eau mesurée ponctuellement en nos deux villes (où la température en valeur discrète ne passe pas nécessairement par 4°C), une probabilité de renversement pour chaque ville a été définie sous la forme d'une gaussienne centrée sur 4°C pour Hawkesbury, et 5°C pour Oka (même si elle est peu mesurée, la température de l'eau brute est d'environ 1°C de moins que l'eau filtrée, selon le chef opérateur de la station). Ces probabilités sont d'amplitude 1 et d'écart-type 0,75°C. Les équations et les courbes associées sont indiquées ci-dessous :

$$Normal\_HAW = \exp\left[-\frac{1}{2}\left(\frac{TEMB\_HAW - 4}{0{,}75}\right)^2\right]$$

$$Normal\_OKA = \exp\left[-\frac{1}{2}\left(\frac{TEMF\_OKA - 5}{0{,}75}\right)^2\right]$$

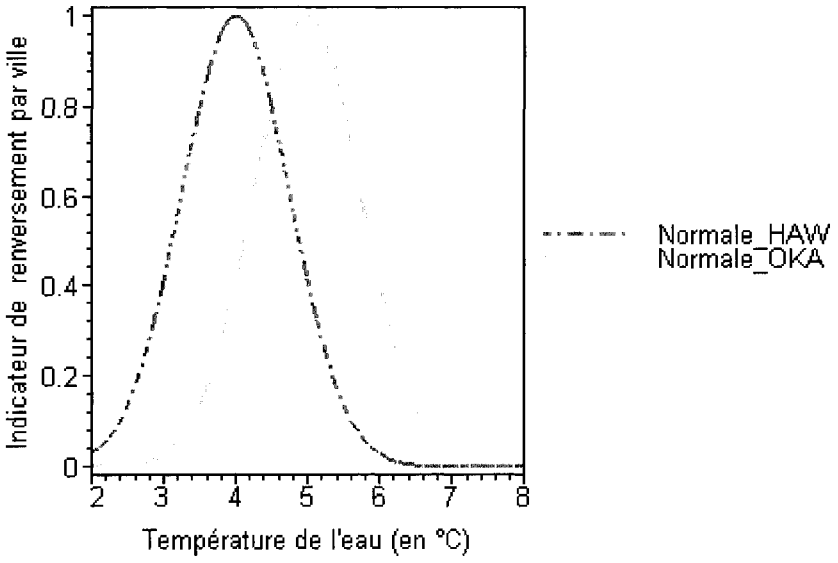

Figure A-1: Probabilité de renversement par ville en fonction de la température de l'eau

On vérifie que les deux probabilités concordent en définissant la variable d'index par :

$$IDX\_RENV = \left[\frac{Normal\_HAW}{2} + \frac{Normal\_OKA}{2} \right]^2$$

Le facteur ½ permet de normaliser la sortie IDX_RENV à 1 et assurer une égale contribution des deux indicateurs. Le carré assure une forte intensité lorsque les pointes concordent, les pointes sont plus francs et brusque ce qui permet une meilleure localisation temporelle estimée du renversement.

Dans la situation où des données de température de l'eau pour l'une des deux stations seraient absentes, l'index est calculé seulement sur un point de mesure, en supposant que Normal_Station $_i$ = 1 (i étant la station dont la mesure est indisponible).

Pourcentage de contribution des Outaouais OUT_FLV

Le but est ici de représenter la contribution de la rivière des Outaouais par rapport au fleuve Saint-Laurent. Cette rivière est bien souvent de bien moindre qualité (microbiologique et physico-chimique) : les opérateurs de la station de Pointe-Claire relatent même un jaunissement net de l'eau brute au printemps lorsque la contribution des Outaouais devient maximale avec le déversement du barrage de Carillon.

Sous l'hypothèse que la prise d'eau de Montréal est située suffisamment loin de la berge pour ne pas être sous influence de la langue d'eau passant par Sainte-Anne de Bellevue (voir la Figure 1-3 de la revue de littérature), la contribution des Outaouais ne tient seulement compte de l'eau passant par le chenal de Vaudreuil.

La contribution des Outaouais sera alors donnée par la formule :

$$OUT\_FLV = \frac{OUT\_VAUD}{(FLV\_CED + FLV\_BEAU)}$$

où OUT_VAUD, FLV_CED, et FLV_BEAU représentent respectivement les débits des Outaouais à Vaudreuil, et du fleuve par Des Cèdres et le barrage de Beauharnois.

Or, cette dernière variable est du domaine du privé (donnée d'Hydro-Québec), elle n'a pas été facile à obtenir dès le début du projet. L'expression fut simplifiée en supposant que les contributions externes entre Beauharnois et Lasalle sont négligeables par rapport au débit du fleuve à Lasalle (FLV_LSL).

D'où la nouvelle expression de la variable OUT_FLV ayant servie pour les calculs :

$$OUT\_FLV = \frac{OUT\_VAUD}{(FLV\_LSL - OUT\_VAUD - OUT\_SAB)}$$

Précipitation maximale journalière PRECX_DS

La réduction du nombre de variables en entrée du réseau tout en gardant le plus d'informations possibles peut se faire en regroupant les précipitations moyennes journalières autour de l'île de Montréal, soit à Sainte-Anne de Bellevue et à Dorval en une variable commune maximum des deux.

PRECX_DS = max (DOR_PREC ; SAB_PREC)

Les précipitations sont des phénomènes très localisés, cependant cette variable testera quelle pourrait être la pluie maximale autour du lac Saint-Louis, pluie influençant le ruissellement et les possibles surverses d'égout sur le lac en amont de la prise d'eau.

Annexe B Courbes et tableaux - turbidité de l'eau brute à Des Baillets

Observation graphique des évènements turbides

Tableau A-1 : Nombre de fois où les facteurs explicatifs ont été activés

	Fin automne			Printemps			Été			Début automne			Automne		
	F	S	I	F	S	I	F	S	I	F	S	I	F	S	I
Fonte des neiges				7	20	6									
Fragilité du couvert	1		5	1	2	1									3
Renversement				7	13	4							2	3	12
Pluie	1	1	5	8	22	26	2	7	20			15	5	7	34
Vent		1	6	8	23	18			11			14	4	8	45
Hausse des tributaires	1	1	4	10	21	12	1	7	8			3	3	3	14
Contribution des Outaouais			8	12	36	26	1	5	1			5	3	5	21
Aucune explication									2			1			
# d'évènements	1	1	10	14	16	30	2	7	27	0	0	19	5	9	52

Tableau A-2 : Pourcentage d'occurrence des facteurs explicatifs des évènements turbides

	Fin automne			Printemps			Été			Début automne			Automne		
	F	S	I	F	S	I	F	S	I	F	S	I	F	S	I
Fonte des neiges			50	50	**56**	20									
Fragilité du couvert	100		50	7	6	3									6
Renversement				50	36	13							40	33	23
Pluie	100	100	50	**57**	**61**	**87**	100	100	74			79	100	78	65
Vent		100	60	**57**	64	60			41			74	80	89	87
Hausse des tributaires	100	100	40	71	**58**	40	50	100	30			16	60	33	27
Contribution des Outaouais			80	86	100	87	50	71	4			26	60	56	40
Aucune explication									7			5			
# d'évènements	1	1	10	14	16	30	2	7	27	0	0	19	5	9	52

Tableau A-3 : Commentaires des résultats de l'analyse des entrées par réseaux PNN

Saison	Seuil (UTN)	Commentaire(s)
Automne		Cette saison est marquée par la prédominance des vents (SAB_VITM-1 et LSF_VITM-1, et avec une moindre influence DOR_VITX-2 et LSF_VITX-2) actifs sur tous les seuils, ils ne seront donc pas rappelés ci-après.
	4	La qualité de l'eau à court terme (TURB_DB-1) semble influencer la prédiction. L'index de renversement et les précipitations sont éliminés rapidement.
	5,5	De même, la qualité de l'eau à très court terme est privilégiée. Les précipitations, l'index de renversement, puis les tributaires s'effacent dans cet ordre.
	7,5	Seuls les vents subsistent au final. Cependant, il est intéressant de noter que les précipitations persistent un peu (notamment DOR_PREC-2 et LSF_PREC-2), ainsi que TURB_HAW-1 et RIV_RAIS-1 qui tiennent jusqu'aux dernières étapes. Précipitations et hausse des tributaires seraient liées.
	9,3	Les descripteurs de qualité, de débit, et de pluie de moyen terme (trois jours et plus) sont éliminés au début. Puis, la qualité, même à Des Baillets, se voit supprimée, pour ne conserver que le vent et les pluies proches (DOR_PREC-2).
Printemps	4	L'index de fonte ressort en premier (surtout autour du 3^{e} jour de décalage), suivi de la qualité de l'eau en amont et les jours précédents à Des Baillets (-3 j à Hawkesbury et -1 j à DB), puis des tributaires secondaires (RAIS-6 en premier) et du vent au final. Les précipitations sont rapidement éliminées.
	5,5	Presque la même chose que 4UTN. L'index de fonte avec un décalage de deux jours ressort, puis qualité de l'eau (TURB_DB-1), tributaires et vents en dernier. Les débits des tributaires à moyen et long terme (RIV_CHAT-3 et -5) sont fortement présents, ils symboliseraient aussi la fonte des neiges. La précipitation à courte terme autour de l'île de Montréal (PRECX_DS-1) persiste un moment avant d'être éliminée.
	7,5	La fonte domine (IDX_FONT-1 ou -2), suivie de la qualité à court et moyen terme (TURB_DB-1 et -3, puis TURB_HAW-3). Les débits restent de bons descripteurs (RIV_RAIS-6 et RIV_CHAT-1 et -3). La contribution des Outaouais (OUT_FLV-1) semble prendre de l'importance ici. Les vents (LSF_VITM-1 et SAB_VITM-1) apparaissent quelques fois. L'index de renversement (IDX_Renv-1) ne ressort que pour une seule des deux répartitions.
	9,3	IDX_FONT-1 reste toujours en tête suivi de très près par les débits de moyen et long terme (RIV_RAIS-6, RIV_CHAT-3 et -5), mais aussi par la qualité 'de proximité' (TURB_DB-1). OUT_FLV-1 confirme sa place pour les événements de haute turbidité. Les conditions météorologiques au lac St François restent présentes (LSF_PREC-2 et LSF_VITM-1).
Été		TURB_DB-1 se distingue en premier, il peut être accompagné de TURB_DB-3; par ailleurs, les tributaires à court terme (1 et 3 jours) apparaissent vite.

Résultats des modèles neuronaux

Figurent ci-après les résultats pour les ensembles Test des répartitions x98 pour les saisons « Printemps », « Eté », et « Automne ».

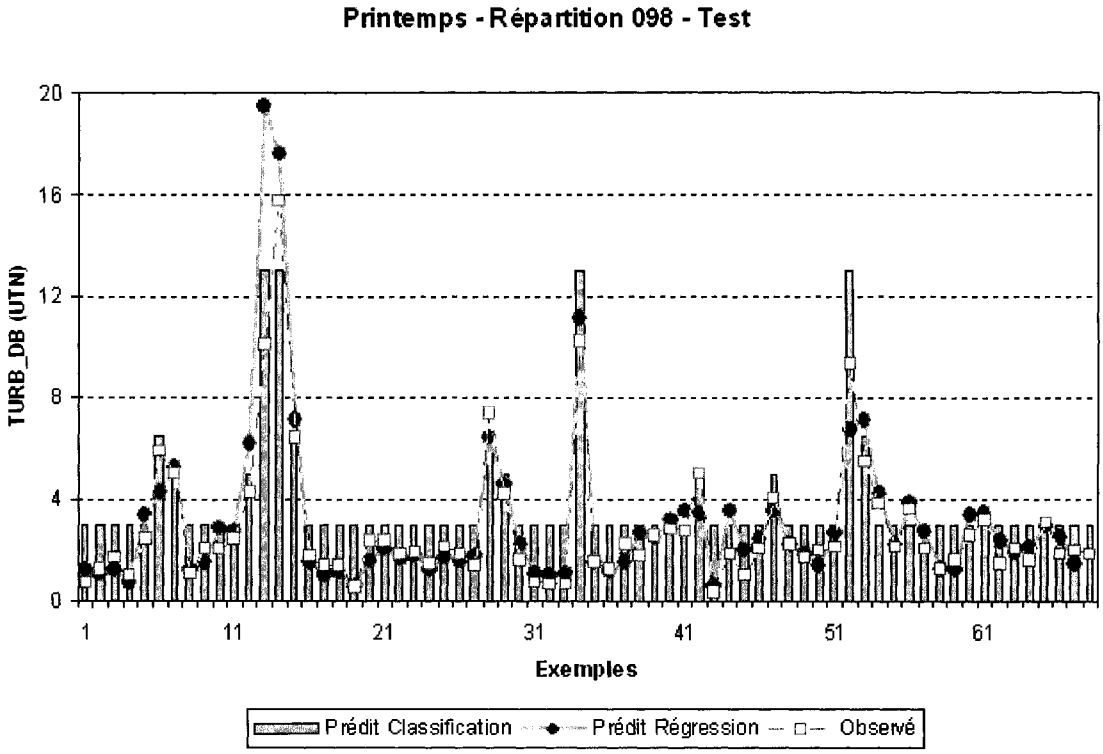

Figure A-2 : TURB_DB observée et prédite par classification et régression au printemps - r098 – Test

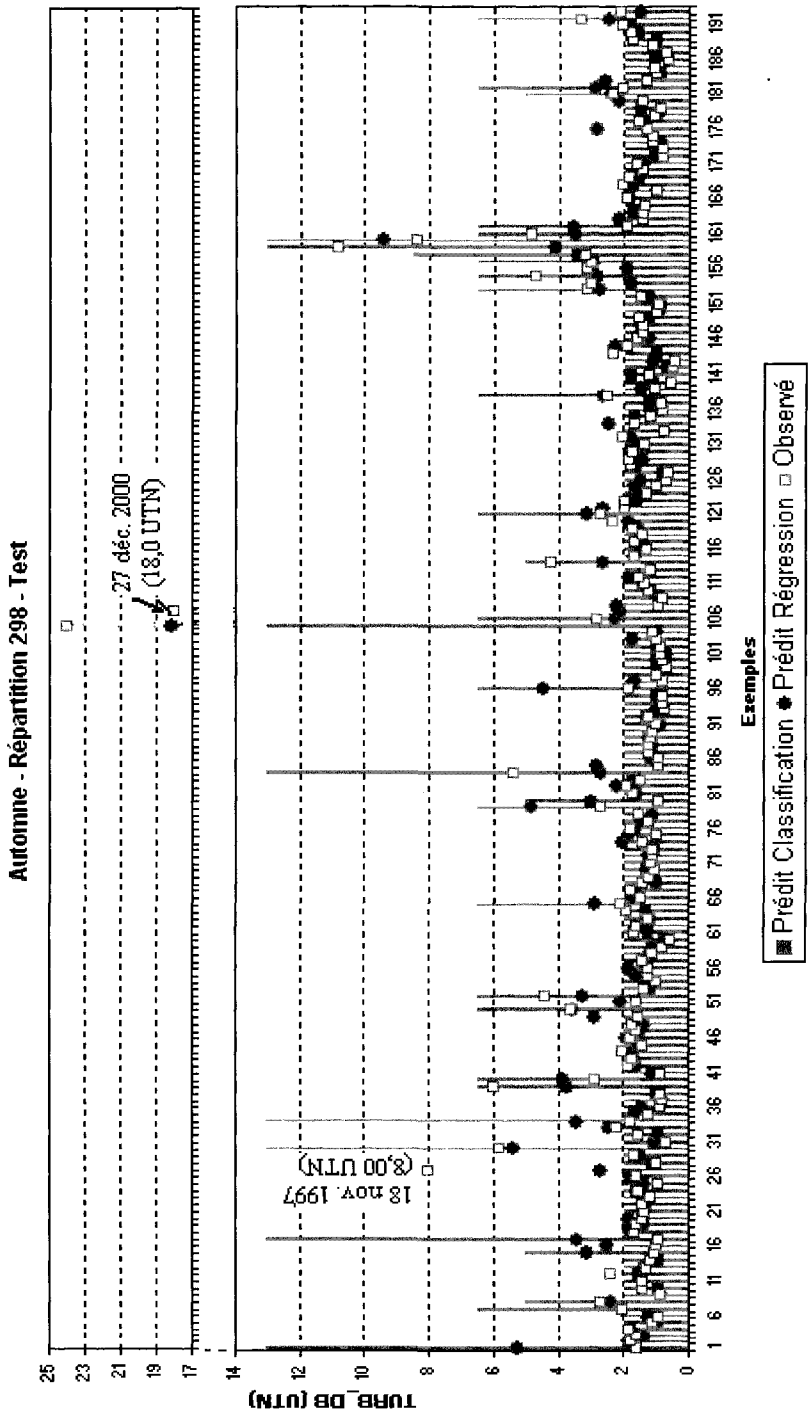

Figure A-3 : TURB_DB observée et prédite par classification et régression à l'automne – r298 – Test

À l'automne, pour la répartition 298, seuls deux évènements de haute turbidité n'ont pas été prédits correctement. Le premier datant du 18 novembre 1997 (8,00 UTN), aurait pour causes potentielles les pluies à Dorval quatre jours avant, ou bien autour du lac Saint-François entre trois et quatre jours avant, selon l'inspection visuelle des évènements turbides. De plus, il semblerait que ce pic isolé soit au début d'un épisode de renversement : l'index de renversement du jour en question est de 80 % et de 40% la veille. Pour cette année nous ne disposions pas des données d'Oka, IDX_RENV étant calculé seulement avec les données d'Hawkesbury. Notons que les modèles classifiant à 7,5 et 9,3 UTN annonce un évènement de haute turbidité ; mais cette prédiction ne ressort pas car les seuils précédents n'étant pas activés, la sortie du modèle classifiant reste à 2 UTN.

Pour le deuxième pic non prédit, celui du 27 décembre 2000 à 18 UTN, nous n'avions trouvé aucune variable explicative activée lors de la phase d'analyse graphique.

Annexe C Partitionnement des exemples

Nomenclature adoptée

Que cela soit pour les groupes d'entrées ou les répartitions, une nomenclature à trois chiffres fut adoptée. Le premier chiffre, celui des centaines, défini la saison ; le chiffre des dizaines défini le seuil de turbidité, et celui des unités est un compteur incrémental au besoin. La nomenclature est résumée dans le tableau suivant :

Tableau A-4 : Nomenclature adoptée pour les répartitions et les groupes d'entrées testés

Chiffre	**Centaine**	**Dizaine**	**Unité**
Signification	Saison	Seuil de turbidité	
Code	« 0 » : Printemps « 1 » : Été « 2 » : Automne	« 0 » : 4 UTN « 1 » : 5,5 UTN « 2 » : 7,5 UTN « 3 » : 9,3 UTN	Aucune contrainte

Remarque : la répartition x01 représente celle choisie avec les années fixes. Elle reste valable pour tous les seuils de turbidité considérés.

Remarque : les répartitions x98 et x99 représentent des échantillonnages aléatoires valables pour tous les seuils de turbidité.

Exemple : « Répartition 211 » signifie la répartition des données d'automne, pour le seuil 5,5 UTN, et il s'agit de la première répartition adoptée pour ce seuil et cette saison.

Méthodes d'échantillonnage spécifique à chaque saison

Découpage par année (x01)

Le découpage par années entières est le plus facile à réaliser. Sur les dix années de données, huit années sont dédiées pour l'apprentissage, un an entier pour Select et l'année restante pour Test. Cette répartition présente l'inconvénient d'être dépendante des conditions à chaque année : beaucoup d'évènements turbides, année de sécheresse, etc. Il est donc plus difficile d'avoir des ensembles représentatifs de la même population. De plus, il se peut qu'un nombre faible d'exemples aux seuils élevés de turbidité soit fréquent (seulement un exemple supérieur à 9,3 UTN au printemps 2005). Ceci motive le recours à la deuxième méthode de découpage : par répartition aléatoire.

Découpage aléatoirement spécifique à chaque seuil de turbidité

Une proportion fixe des exemples bas puis hauts est allouée aléatoirement dans chaque ensemble. Le nombre d'exemples diminuant avec l'augmentation de la valeur seuil de turbidité, une répartition propre à chaque seuil est donc créée. Le pourcentage des ensembles Select et Test varie de 10 à 20 % dépendamment du nombre total d'exemples hauts disponibles. Chacune de ces répartitions est spécifique à sa saison et à son seuil.

Découpage aléatoire valable pour tous les seuils (x98 et x99)

La répartition précédente donne de bons résultats en termes de similitudes des ensembles, cependant, lors de la construction du modèle final et la mise en commun de tous les modèles, un échantillonnage efficace pour tous les seuils est indispensable. Les exemples en deçà de 4 UTN sont répartis comme ceux des répartitions x02 et x11 (seuils 4 et 5,5 UTN respectivement) ; puis, les données sont triées par turbidité croissante (TURB_DB). Entre chaque valeur seuil de turbidité, entre 8 et 16 % des exemples dans sont répartis dans Select et Test en respectant les quatre consignes suivantes :

- Toujours mettre les exemples maximaux de turbidité dans Train.
- Garantir un minimum de deux exemples hauts dans chaque ensemble.
- Utiliser des années différentes dans chaque ensemble (exemple : Test ne doit pas contenir uniquement des exemples de 1998).
- Ne pas concentrer les exemples d'un ensemble dans un intervalle de turbidité restreint (exemple : dans l'intervalle [4 ; 5,5 UTN], il ne faut pas que ces trois exemples de turbidité soit extraits de l'intervalle [4,0 ; 4,2 UTN]).

Tableau récapitulatif – turbidité à l'eau brute de Des Baillets

Dans les tableaux ci-après, figurent les détails de construction de chaque répartition. Dans le tableau suivant, pour les échantillonnages issus d'un découpage année par année, sont indiquées les années des ensembles Select et Test ; et, dans le cas des échantillonnages aléatoires, la proportion des exemples allouée dans chaque ensemble.

Tableau A-5 : Détails de construction de chaque répartition

Répartition #	Saison	Seuil de turbidité (en UTN)	Année ou % SELECT	Année ou % TEST
1	Printemps	Tous	1997	2005
2		4	10%	10%
11		5,5	10%	10%
21		7,5 et 9,3	15%	15%
98		Tous	10,60%	10,90%
99		Tous	8,60%	10,90%
101	Été	4	1999	2000
102		4	20%	20%
201	Automne	Tous	2000	2003
202		4	15%	15%
211		5,5	15%	15%
221		7,5	15%	15%
231		9,3	15%	15%
298		Tous	15,30%	15,90%
299		Tous	15,60%	14,50%

Les exemples de chaque répartition sont répartis comme indiqué dans le tableau suivant.

Tableau A-6 : Nombre d'exemples disponibles par répartitions

PRINTEMPS

		Répartition #	001			002	011	021	021	098			099		
		Seuil (UTN)	4	5,5	7,5	9,3	4	5,5	7,5	9,3	4	5,5	7,5	9,3	4	5,5	7,5	9,3
	Turbidité observée																	
Ensemble	LEARN	Basse	434	472	485	488	418	463	429	436	412	450	462	469	418	459	473	480
		Haute	68	31	19	16	74	33	20	13	70	32	20	13	77	36	22	15
	SELECT	Basse	43	50	52	54	44	55	73	73	57	63	63	65	45	49	50	51
		Haute	13	5	4	2	10	7	3	3	10	4	4	2	8	4	3	2
	TEST	Basse	44	48	51	55	59	52	86	88	54	59	63	63	60	64	65	66
		Haute	12	8	5	1	9	4	5	3	13	8	4	4	8	4	3	2

ETE

		Répartition #	101	102
		Seuil (UTN)	4	4
	Turbidité observée			
Ensemble	LEARN	Basse	1057	822
		Haute	11	9
	SELECT	Basse	132	247
		Haute	1	2
	TEST	Basse	131	251
		Haute	2	3

AUTOMNE

		Répartition #	201			202	211	221	231	298			299		
		Seuil (UTN)	4	5,5	7,5	9,3	4	5,5	7,5	9,3	4	5,5	7,5	9,3	4	5,5	7,5	9,3
	Turbidité observée																	
Ensemble	LEARN	Basse	997	1022	1043	1050	855	871	911	896	840	868	887	893	855	883	902	908
		Haute	65	40	19	12	61	42	19	14	68	40	21	15	67	39	20	14
	SELECT	Basse	121	125	127	127	196	196	190	210	192	197	200	201	196	201	204	205
		Haute	9	5	3	3	13	7	5	3	11	6	3	2	11	6	3	2
	TEST	Basse	113	122	124	126	180	202	193	197	198	203	206	208	180	185	188	190
		Haute	18	9	7	5	18	5	5	3	12	7	4	2	13	8	5	3

Annexe D Prétraitement des entrées du modèle

Dans le choix des variables candidates retenues pour l'élaboration des réseaux de neurones, il convient d'investiguer l'influence du prétraitement sur la performance obtenue par le réseau. Dans un premier temps, la fonction du prétraitement sera rappelée, ensuite deux prétraitements seront considérés : min-max et fonction de répartition couplée à un min max.

Pourquoi pré traiter les variables ?

Ceci n'est pas indispensable et il serait possible de laisser les variables telles quelles. Le réseau, lors de sa phase d'apprentissage, peut jouer sur ces poids et biais pour identifier l'importance relative de chaque variable sur la sortie. Cependant ce processus peut être assez long car il implique avant l'apprentissage du phénomène à modéliser de se concentrer sur l'étalonnage des valeurs numériques des variables. Ceci entraîne un double problème.

Tout d'abord, il semble évident que toutes les valeurs numériques ne traduisent pas l'importance relative des entrées sur la sortie à modéliser. Considérons l'exemple suivant : modélisation la turbidité du jour J en fonction de la turbidité et de la conductivité de la veille. Le rapport entre les valeurs de turbidité et de conductivité est de l'ordre de 10 à 100, ainsi tant que les poids ne sont pas correctement étalonnés, les valeurs de turbidité peuvent être considérées négligeables par rapport à la conductivité, bien que la variable soit importante pour la prédiction.

De plus, des variables brutes avec de fortes valeurs numériques, qui ne sont pas pondérées par de faibles poids des neurones (ce qui est le cas au début de l'apprentissage où les poids sont choisis aléatoirement), peuvent entraîner une forte valeur en entrée de la fonction d'activation, ici tangente hyperbolique dans la couche cachée. Ainsi, si la fonction d'activation démarre dans sa zone extrême où sa dérivée

est quasiment nulle, la dérivée relative de la sortie du neurone par rapport aux poids sera faible également. Donc, le gradient de l'erreur commise par rapport aux poids des connexions sera lui aussi très faible. Lors de la phase d'apprentissage par l'algorithme de rétropropagation, l'avancée dans l'espace des poids est proportionnelle au gradient de l'erreur commise. Ceci entraînerait donc des déplacements minimes dans l'espace des poids lors de l'étalonnage de ces derniers, d'où une phase d'apprentissage très longue.

Pour plus de détails sur le prétraitement, se reporter au chapitre 8 du livre de Bishop (1995). En conclusion de ce paragraphe, il est important de retenir que le prétraitement doit ramener les valeurs numériques des variables dans la zone de fonctionnement linéaire des fonctions d'activation (Figure A-4). Les variables seront centrée-réduites dans la plage [-0,8 ; 0,8]. La sortie du modèle de régression sera post-traitée pour coder les valeurs numériques dans la plage de sortie de la fonction d'activation, ici pour une sigmoïde [0;1]. Cette fonction doit impérativement être réversible.

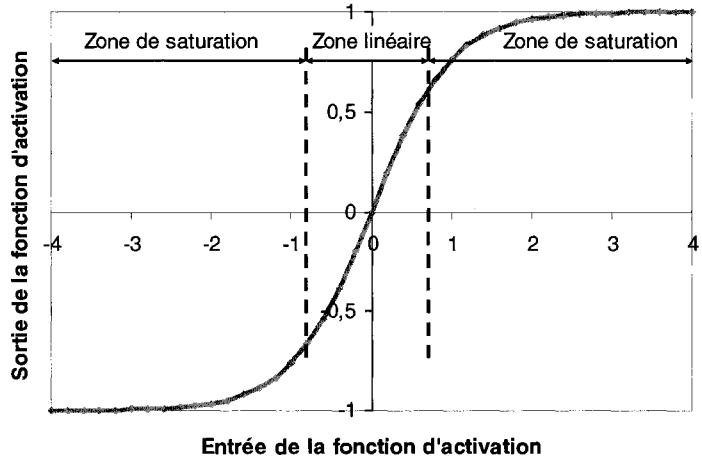

Figure A-4 : Fonction d'activation tangente hyperbolique

Différents prétraitements

Après avoir vu l'importance du prétraitement afin d'aider le réseau dans son apprentissage, voici les deux prétraitements différents considérée pour l'étude : min-max et la fonction de répartition couplée à une transformation linéaire entre [-0,8 ; 0,8].

Commentaire sur le choix du prétraitement avec fonction de répartition

La connaissance préalable de la fonction à modéliser est un élément clef de l'optimisation de la performance du réseau (Haykin, 1999). Le théorème de Bayes définit que la probabilité qu'un exemple **x** puisse être classé dans la classe $(C_k)_{k\in\{basse;haute\}}$ s'écrit :

$$P(C_k|\mathbf{x}) = \frac{P(\mathbf{x}|C_k)P(C_k)}{P(\mathbf{x})}$$

La probabilité a posteriori $P(C_k|\mathbf{x})$ est fonction de la probabilité de vraisemblance ($P(\mathbf{x}|C_k)$, déterminée lors de la calibration du réseau, de la probabilité a priori ($P(C_k)$) et d'un facteur de normalisation ($P(\mathbf{x})$) indépendant des classes (Bishop, 1995). Il semble clair qu'il soit important d'avoir une connaissance préalable sur la répartition des données : $P(C_k)$ ou $P(\mathbf{x})$. Ceci afin de guider le réseau dans son apprentissage en le forçant par des hypothèses préalables sur les variables.

Fonctionnement des transformations

La fonction de répartition associe à toute valeur x numérique d'une variable le pourcentage du nombre de mesures inférieures ou égales à x. C'est donc une fonction bijective et monotone de $\Re$ vers [0; 1]. Ainsi, à toute valeur de la variable d'entrée correspond une et une seule valeur de sortie comprise entre zéro et un. Cette

transformation est donc réversible. Il convient dans un premier temps de définir quelle distribution convient le mieux aux variables (ceci fait l'objet de l'Annexe E, ci-après). Ensuite, les prétraitements peuvent être menés.

Voici l'exemple pour la turbidité à l'automne dans le cas où l'objectif serait de transformer les données de $\Re$ vers [0; 1]. À partir des données brutes, il ressort que la distribution log-normale convient le mieux à la variable TURB_DB-1 (Figure A-5 a). En abscisse, la valeur de la turbidité en UTN, en ordonnées le nombre d'observations. Nous en déduisons sa fonction de répartition (Figure A-5 b) en trait plein. En ordonnée, la fréquence cumulative en pourcentage donnera le résultat du prétraitement par fonction de répartition. En trait d'axe, est tracé la droite passant par les points [Turb_DB-1_{min}; 0%] et [Turb_DB-1_{Max}; 100%], cette droite donnera le résultat de la transformation min-max.

Illustration numérique : sur le schéma b), à la valeur Turb_DB-1 = 4UTN seront associées les valeurs prétraitées 0,18 et 0,84, pour les méthodes min-max et fonction de répartition respectivement.

Un calcul identique est effectué pour chaque variable de chaque saison. Si l'objectif était de ramener les données prétraitées dans la plage [-0,8 ; 0,8], il suffirait de rajouter une transformation linéaire de [0; 1] vers [-0,8 ; 0,8] en dernière étape.

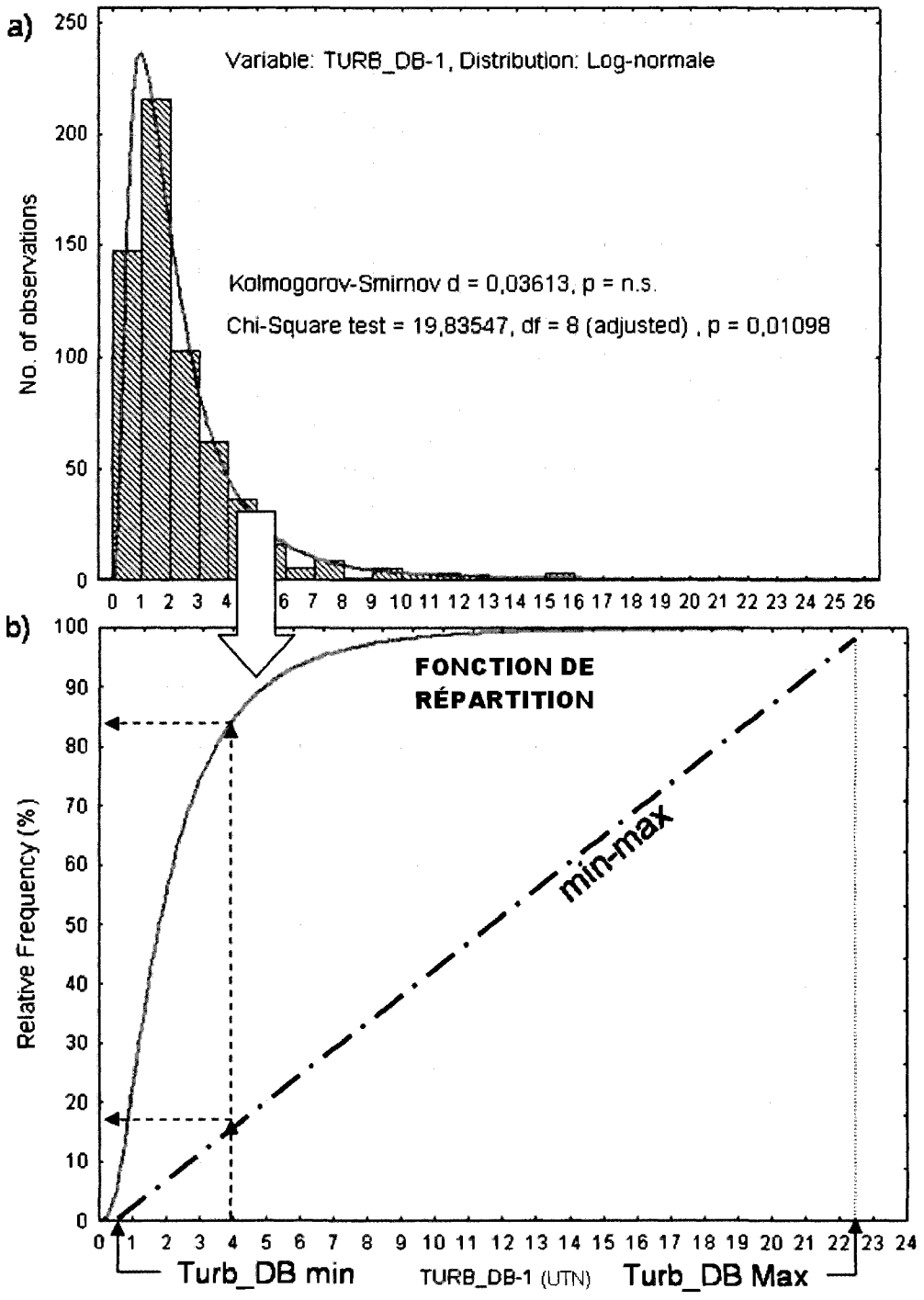
a)
Variable: TURB_DB-1, Distribution: Log-normale
Kolmogorov-Smirnov d = 0,03613, p = n.s.
Chi-Square test = 19,83547, df = 8 (adjusted) , p = 0,01098
No. of observations
b)
FONCTION DE RÉPARTITION
min-max
Relative Frequency (%)
Turb_DB min
TURB_DB-1 (UTN)
Turb_DB Max

Figure A-5 : (a) Distribution log-normale associée à la TURB_DB-1 à l'automne. (b) Fonction de répartition associée à la loi log normale et transformation linéaire min-max

Annexe E Distributions adoptées pour les variables

Comme nous l'avons vu dans l'annexe précédente, afin de réaliser un prétraitement spécifique à chaque variable, il faut connaître les distributions de ces dernières. Cette annexe décrit la méthodologie et le matériel utilisé à cette fin.

Matériel et méthode

Deux logiciels furent utilisés. Tout d'abord l'extension d'Excel® appelée Crystal Ball® afin d'effectuer une recherche globale par variable et par saison de la meilleure distribution parmi celles disponibles dans le logiciel. À savoir : gaussienne, log-normale, bêta, gamma, weibull, exponentielle. Le logiciel sort un classement des distributions selon trois critères de test : le test du khi deux, celui de Kolmogorov-Smirnov (K-S), et celui d'Anderson Darling. L'équation et les paramètres internes de la meilleure distribution sont ensuite rentrés sous Statistica® afin de modifier la base de données. Cette opération est effectuée pour toutes les variables retenues dans chaque saison sauf les variables d'index car ces dernières furent déjà construites pour être comprises entre zéro et un.

Critère de sélection d'une distribution

Nous avons choisi de privilégier les tests du K-S et celui d'Anderson Darling au khi deux. Ce dernier présente l'inconvénient d'être sensible au nombre d'intervalles avec lesquels on discrétise notre variable (i.e. le nombre de barres sur l'histogramme, voir Figure A-5a).

La valeur de sortie du test K-S représente l'écart absolu maximal entre la fonction de distribution cumulée observée et celle prédite. On appelle valeur critique du K-S la valeur en deçà de laquelle nous pouvons accepter la distribution considérée pour modéliser la série de données. La valeur critique du K-S ne dépend pas la distribution adoptée, mais seulement du risque de rejeter faussement l'hypothèse « H_0 : aucune

différence entre les distributions » (généralement 5%), et du nombre d'exemples disponibles (ce qui affecte la résolution de la courbe de distribution cumulée observée).

Le test d'Anderson Darling est une variante du test du K-S qui met l'accent sur l'extrémité des distributions, ses valeurs critiques dépendent en plus de la distribution considérée (NIST/SEMATECH, 2006).

Comme nous nous intéressons aux phénomènes extrêmes nous avons privilégié le test d'Anderson Darling, puis le test du K-S, et finalement regarder la valeur du khi deux pour s'assurer qu'il y ait une bonne approximation globale de la distribution (les valeurs des deux premiers tests étant des indicateurs ponctuels). Il est rare que nous obtenions une valeur de K-S inférieure à la valeur critique, cependant elle s'en approche pour la plupart des variables excepté la pluie. Un examen graphique complémentaire permet de dire si la distribution convient. Rappelons-le le but n'est pas de trouver la meilleure distribution, mais d'extraire une tendance générale quant à la répartition de nos données, ainsi deux distribution aux valeurs de K-S voisines pourraient convenir.

Tableau récapitulatif des distributions par saison

Voici ci-après un tableau récapitulatif par saison des distributions retenues pour les variables. Chaque variable n'a été considérée qu'une seule fois car nous avons émis l'hypothèse qu'un décalage temporel de quelques jours n'aurait pas d'impact majeur sur la distribution finale ; en effet, le changement de quelques valeurs sur les centaines présentes pour le calcul ne doit pas changer fondamentalement nos résultats. À titre indicatif, sont données les valeurs du test du K-S et la valeur critique associée pour une probabilité de se tromper de 5%. Celle-ci est calculée par la formule :

$1{,}36 / \sqrt{\text{Nombre d'exemples}}$ (si l'échantillon est plus grand que 35 exemples).

Tableau A-7: Récapitulatif des distributions retenues - automne

Saison	Variable	Moyenne	Ecart type	Distribution retenue	K-S	Paramètres internes
Automne	TURB_DB-1	1,85	1,25	Lognormal	0,096	
	TURB_HAW-1	4,01	2,91	Lognormal	0,047	
	RIV_RAIS-1	3,18	5,56	Weibull	0,106	Location=0,00;Scale=2,13;Shape=0,60308
	RIV_BAUD-1	1,07	1,61	Weibull	0,056	Location=0,00;Scale=0,83;Shape=0,68243
	RIV_CHAT-1	25,98	26,82	Lognormal	0,067	
	DOR_VITM-2	16,20	6,94	Beta	**0,014**	Minimum=2,5;Maximum=90,6;Alpha=3,12437;Beta=16,99919
	LSF_VITM-1	12,01	5,66	Beta	**0,018**	Minimum=1,8;Maximum=49,9;Alpha=2,32578;Beta=8,66364
	LSF_VITX-2	22,02	8,47	Beta	0,045	Minimum=-2;Maximum=302;Alpha=7,45046;Beta=86,2358
	SAB_VITM-1	12,48	5,36	Beta	**0,019**	Minimum=1,7;Maximum=53,2;Alpha=3,00769;Beta=11,32047
	DOR_PREC-2	2,88	23,61	Gamma	0,298	Location=-0,1;Scale=7,9;Shape=0,37942
	LSF_PREC-2	2,66	4,60	Gamma	0,332	Location=-0,1;Scale=7,7;Shape=0,35634
	PRECX_DS-1	3,44	5,70	Gamma	0,250	Location=-0,1;Scale=9,1;Shape=0,39091
# d'exemples	1322					
Valeur critique du K-S	0,037					

* en gras figurent les valeurs de K-S inférieures ou égales à la valeur critique

Tableau A-8 : Récapitulatif des distributions retenues - printemps

Saison	Variable	Moyenne	Ecart type	Distribution retenue	K-S	Paramètres internes
Printemps	TURB_DB-1	2,45	2,26	Lognormal	**0,036**	Mean=2,45;Std. Dev.=2,26
	COUL_DB-1	9,65	4,13	lognormal	0,074	Mean=9,65;Std. Dev.=4,13
	TURB_HAW-2	9,74	9,75	Lognormal	0,067	Mean=9,74;Std. Dev.=9,75
	OUT_FLV-1	8,49	4,69	Weibull	**0,042**	Location=1,76;Scale=7,42;Shape=1,45799
	RIV_RAIS-6	18,01	28,42	Lognormal	0,063	Mean=18,01;Std. Dev.=28,42
	RIV_BAUD-4	6,66	10,55	Lognormal	**0,050**	Mean=6,66;Std. Dev.=10,55
	RIV_CHAT-1	89,36	100,46	Lognormal	**0,052**	Mean=89,36;Std. Dev.=100,46
	DOR_VITM-2	16,88	6,78	Beta	**0,020**	Minimum=2,1;Maximum=110,1;Alpha=3,95173;Beta=24,95459
	DOR_VITX-1	29,19	9,63	Gamma	0,058	Scale=3,15386896;Shape=9,25527835
	LSF_VITM-1	13,06	5,69	Beta	**0,023**	Minimum=2,1;Maximum=186,8;Alpha=3,44486;Beta=54,53716
	LSF_VITX-2	23,74	8,22	Gamma	**0,046**	Scale=2,76044075; Shape=8,60017000
	SAB_VITM-1	13,40	5,03	Weibull	**0,030**	Location=4,1;Scale=10,5;Shape=1,93149
	DOR_PREC-2	2,45	4,56	Gamma	0,377	Location=-0,1;Scale=8,3;Shape=0,30028
	LSF_PREC-2	2,21	4,00	Gamma	0,365	Location=-0,1;Scale=7,1;Shape=0,32086
	PRECX_DS-5	3,48	6,09	Gamma	0,291	Location=-0,1;Scale=10,4;Shape=0,34131

# d'exemples	Valeur critique du K-S
616	0,055

* en gras figurent les valeurs de K-S inférieures ou égales à la valeur critique

Tableau A-9: Récapitulatif des distributions retenues - été

Saison	# d'exemples	Valeur critique du K-S	Variable	Moyenne	Ecart type	Distribution retenue	K-S	Paramètres internes
Eté	1385	0,037	TURB_DB-1	2,11	0,61	Gamma	0,0477	Location=0,37;Scale=0,21;Shape=8,20226
			COUL_DB-1	7,11	2,55	Gamma	0,1463	Location=2,81;Scale=1,51;Shape=2,85056
			RIV_RAIS-1	5,03	42,53	Lognormal	0,0564	
			RIV_BAUD-1	1,08	5,27	Lognormal	**0,029**	
			RIV_CHAT-1	23,79	31,93	Lognormal	0,0515	
			DOR_VITX-2	25,98	8,16	Beta	0,0675	Minimum=8;Maximum=98;Alpha=3,82072;Beta=15,04479
			LSF_VITM-1	9,69	4,21	Beta	**0,0194**	Minimum=-0,1;Maximum=133,0;Alpha=4,88819;Beta=61,81083
			LSF_VITX-2	19,03	6,62	Gamma	0,0477	Location=-7;Scale=2;Shape=15,758
			SAB_VITM-1	10,92	4,14	Beta	**0,0226**	Minimum=3,0;Maximum=33,1;Alpha=2,40816;Beta=6,76031
			DOR_PREC-2	2,66	4,81	Gamma	0,3563	Location=-0,1;Scale=8,4;Shape=0,3284
			LSF_PREC-2	2,82	4,96	Gamma	0,3641	Location=-0,1;Scale=8,4;Shape=0,35303
			PRECX_DS-1	3,48	6,12	Gamma	0,3128	Location=-0,1;Scale=10,5;Shape=0,34045

* en gras figurent les valeurs de K-S inférieures ou égales à la valeur critique

Annexe F Intelligent Problem Solver de Statistica

Afin de déterminer quelle configuration (entrées – prétraitement – et nombre de neurones de la couche cachée) donne le modèle le plus performant, nous avons utilisé l'algorithme de recherche heuristique de Statistica appelé *Intelligent Problem Solver* (IPS). Pour des architectures de réseaux pré définies, soit par exemple le type de réseau (perceptron multicouches, réseau à base radiale, etc), la plage dans laquelle le nombre de neurones des couches cachées va varier, les entrées varient aussi.etc. L'IPS va effectuer l'apprentissage de x réseaux et en retenir les y parmi x meilleurs. Une macro-instruction programmée en Visual Basic®, a permis d'utiliser l'analyse de l'IPS en fixant certains paramètres et en faisant varier les autres. Ces paramètres variables furent notamment : l'ensemble d'entrées utilisé, la répartition utilisée, le nombre de neurones dans la couche cachée et le prétraitement ou non des entrées (par fonction de répartition, voir Annexe D). À chaque réseau testé, les critères de performance pré-établis sortaient un scalaire et l'apprentissage répété du réseau indiqua la variabilité de la performance, donc la stabilité du résultat. Au final, nous pouvions tracer des courbes de type boîtes à moustaches du critère de performance considéré en fonction du nombre de neurones de la couche cachée et identifier le meilleur modèle répondant à nos besoins.

Afin que les expériences puissent être reconduites, nous allons décrire dans cette annexe quels furent les paramètres utilisés.

Organisation des bases de données

Pour chaque saison, nous disposions de deux fichiers de données. Le premier pour les données brutes, le deuxième pour les données pré traitées. Ces fichiers contenaient les variables suivantes :

- Toutes les entrées à tester, brutes ou pré traitées par fonction de répartition.

- Des variables de type texte stockant les classes de turbidité (classe de I à V, ou seuils 4 ; 5,5 ; 7,5 et 9,3 UTN ayant comme sortie « basse » ou « haute »).
- Des variables texte stockant l'échantillonnage utilisé (« Train », « Select » et « Test »).

Algorithme sommaire de la macro-instruction

1. Ouverture du fichier de données spécifié (fixant la saison et le recours au prétraitement ou non).
2. On fixe les variables de seuil de turbidité, d'échantillonnage, d'entrées utilisées et les neurones de la couche cachée à tester.

Exemple : dans la macro-instruction *est écrit*

«Analyse_Main("Seuil4UTN","Repartition102", 103,1,20,1,22,40,2,45,80,5)»

Ceci veut dire que pour le seuil 4UTN, l'échantillonnage numéro 102 et le groupe d'entrées numéro 103, nous allons obtenir les résultats pour la plage de neurones suivante : de 1 à 20 neurones avec un pas de 1, de 22 à 40 neurones avec un pas de 2, et de 45 à 80 neurones avec un pas de 5.

3. Création des tableaux récapitulatifs qui vont contenir tous les résultats de nos expériences.
4. Analyse IPS modifiée pour chaque nombre de neurones de la couche cachée.
5. Calcul des résultats selon les critères de performance choisis.
6. Ecriture des résultats dans les tableaux récapitulatifs appropriés.
7. Répétition des étapes 2 à 6 pour chaque configuration de notre plan d'expérience.

Détails des paramètres de l'analyse IPS modifiée

Cette partie fut rédigée à l'aide de l'aide électronique de Statistica disponible en ligne (Statsoft, 2006).

Rendu à l'étape 4 de la macro-instruction précédente, sont fixés les paramètres suivants : saison, combinaison d'entrées, prétraitement ou non, seuil turbide de classification, répartition des exemples selon la variable d'échantillonnage, nombre de neurones dans la couche cachée. Nous appellerons cette combinaison de notre plan d'expérience, une « *configuration de réseau* ».

Seulement le type de réseau dit perceptron multicouches (PMC) à une seule couche cachée sera considéré (voir la section 2.3.7). Le groupe d'entrées fut fixé : il n'y a pas de détermination de sous-ensemble d'entrées par « *pruning* ».

Architecture

L'IPS effectue un prétraitement automatique linéaire des entrées. Par la méthode du min-max, il ramène automatiquement nos entrées dans la plage [0 ; 1]. C'est pourquoi les fichiers de données ne comprenaient que les données brutes ou transformées par la fonction de répartition. La transformation linéaire finale étant assurée par le logiciel.

Quant aux fonctions d'activation, nous avons choisi la tangente hyperbolique dans la couche cachée et la fonction sigmoïde dans la couche de sortie. Tanh est particulièrement adaptée pour les PMC avec son caractère antisymétrique permettant d'accélérer l'apprentissage (Haykin, 1999), alors que la fonction sigmoïde (bornée entre 0 et 1) peut correspondre à une probabilité d'appartenance à une classe ou l'autre dans le cas de classificateur à un seuil. La valeur 0 étant « turbidité basse », la valeur 1 étant « turbidité haute ». Pour la régression, sigmoïde permet de donner des prédictions physiquement plausibles (pas de valeurs de turbidités prédites négatives)

Dans le problème de classification, le seuil de classification optimal (i.e. scalaire compris entre 0 et 1 définissant la frontière entre basse et haute turbidité) et déterminé par le logiciel lui-même par le tracé de courbes ROC. Les courbes ROC, largement utilisées en santé publique permettent de trouver un compromis optimal entre

sensibilité et spécificité (Moise et al., 1986). L'objectif est de maximiser le classement des évènements hauts tout en minimisant les faux positifs. Dans l'analyse IPS, nous avons choisi un coefficient de perte de deux : sur la courbe ROC Statistica cherche le point où le rapport faux positif sur faux négatif vaut deux, indépendamment du nombre d'exemples dans chaque classe (Statsoft, 2006). Ceci veut dire que les évènements de haute turbidité ont été choisis arbitrairement deux fois plus importants dans l'élaboration des modèles.

Apprentissage

À configuration donnée, l'analyse va effectuer 40 apprentissages de réseaux avec une initialisation des poids différente (choisie aléatoirement à chaque ré-apprentissage selon une distribution uniforme comprise entre 0 et 1). Cette initialisation nous permet de partir d'un point différent de l'espace des poids et de converger vers divers minima locaux, ou globaux si possible, de la surface d'erreur.

L'apprentissage se déroule en deux phases où nous avons laissé les paramètres par défaut du logiciel. La première phase sur 100 époques avec l'algorithme de rétropropagation (taux d'apprentissage $\eta = 0,01$ et *momentum* $\mu = 0,3$). La deuxième phase se faisant avec l'algorithme du gradient conjugué sur 500 époques. Plus de détails sur ces algorithmes peuvent être retrouvés au chapitre 4 du livre de Haykin (1999). L'algorithme de Levenberg-Marquardt étant plus adapté pour les problèmes de régression avec de plus petits réseaux (nombre de neurones cachés inférieurs à 100), il sera utilisé pour le modèle de régression (Statsoft, 2006). Une fois ces époques terminées, Statistica récupère le meilleur réseau obtenu précédemment (minimum de l'erreur considérée sur l'ensemble Select). Cette méthode, dite de validation croisée, permet d'optimiser l'arrêt de l'apprentissage lorsque l'erreur de généralisation augmente et que l'erreur d'apprentissage continu de diminuer, elle permet de prévenir le phénomène du sur-apprentissage. Nous n'avons pas modifié les paramètres

d'apprentissage car la convergence vers un « bon » minima local s'effectuera par le biais de nos deux phases d'apprentissage mêlant méthodes du premier ordre et du deuxième ordre : abaisser le taux d'apprentissage ne conduirait pas à des solutions significativement meilleures, sous réserve que l'erreur de diverge pas (Özesmi et al., 2006). Les principaux facteurs influençant nos résultats sont les entrées utilisées et le nombre de neurones de la couche cachée.

À chaque époque de la phase d'apprentissage les exemples sont mélangés afin de ne pas être soumis au réseau dans un ordre séquentiel fixe. Ceci rend l'algorithme moins susceptible de tomber dans un minimum local de la surface d'erreur.

L'algorithme pratique de plus une méthode appelée « *pruning* » pour éliminer les neurones cachés dont la contribution est quasi nulle, soit où tous les poids en sortie du neurone sont en deçà du seuil 0.05 (valeur par défaut). Par conséquent, mettre plus de neurones que nécessaires tendra à faire stagner les performances car l'IPS éliminera des connexions. Cette option est désactivée pour la recherche des modèles de classification.

Réseaux retenus

De nos 40 réseaux calibrés, nous ne retenons que les 10 meilleurs résultats de l'ensemble de sélection. La fonction d'erreur considérée est l'entropie ou la somme des carrés des erreurs pour les problèmes de classification et de régression respectivement. Les résultats obtenus étant probabilistes, les chiffres arbitraires 40 et 10 réseaux furent choisis pour retenir seulement 25% des réseaux testés parmi un nombre qui puissent être un compromis entre un temps de calcul raisonnable et la détermination non biaisée de l'erreur moyenne d'une configuration de réseau.

Les différents paramètres utilisés lors de l'analyse IPS sont résumés dans le tableau récapitulatif ci-après.

Tableau A-10 : Récapitulatif des paramètres de l'analyse Intelligent Problem Solver modifié

Saison	Fixée par base de donnée (automne, printemps, été)	
Prétraitement	Fixée par base de donnée, puis min-max sur [0;1]	
Type de modèle	**Classification**	**Régression**
Variable de sortie	Variable binaire seuil (« basse »/ « haute »)	Variable continue DIFF_TURB
Variables d'entrée	Fixée dans la macro-instruction (plan d'expérience)	
Détermination d'un sous-ensemble d'entrées	Non	Oui
Échantillonnage	Fixée dans la macro-instruction (plan d'expérience)	
Nombre de neurones cachés	Fixée dans la macro-instruction	
Fonctions d'activation	Couche cachée : tangente hyperbolique Couche de sortie : sigmoïde	
Nombre des réseaux testés	40 réseaux	Spécial : on fait tourner l'IPS pendant deux jours
Fonction d'erreur	Entropie	Somme des carrés de l'erreur
Apprentissage	Maximum de 100 époques de rétropropagation (η=0,01 et μ=0,3) Maximum de 500 époques de gradient conjugué	
Arrêt de l'apprentissage	Validation croisée : arrêt de l'apprentissage lorsque l'erreur de l'ensemble de sélection est minimale	
Méthode de régularisation	Méthode de régularisation des poids de Weigend	
« *Pruning* »	Non	Oui, seuil de sortie 0,05
Seuil de classification optimal	Déterminé automatiquement par courbe ROC. Coefficient de perte = 2	Ne s'applique pas
Nombre de réseaux retenus	10 réseaux donnant le minimum de la fonction d'erreur sur l'ensemble de sélection	

Annexe G Critères de performance retenus

Rappel des objectifs

Après avoir obtenu par simulation toute une série de résultats, il convient de pouvoir les transformer dans un format plus facilement interprétable, représentatif de nos objectifs de modélisation, et uniformisé pour pouvoir comparer différents modèles entre eux.

Nos objectifs de modélisation sont rappelons-le de pouvoir prédire les pointes de turbidité. Les conséquences en termes de traitement peuvent devenir sérieuses à partir de 4UTN à l'eau brute, au-delà la prédiction d'un pic de 9 ou 12 UTN ne va pas être fondamentalement différente pour l'opérateur s'il peut être averti de l'arrivée imminente d'une situation de crise. Ainsi une mauvaise classification dans la branche inférieure à 4UTN (prédit 1UTN alors qu'observé 3UTN) sera moins importante qu'un faux positif à 4 UTN (prédit supérieur à 4UTN alors qu'observé inférieur à 4UTN), qui sera moins importante qu'un faux négatif (prédit inférieur à 4UTN alors qu'observé supérieur à 4UTN). Dans la même idée, avec nos modèles de classificateurs à un seuil implémentés en cascade, un faux négatif empêche le modèle suivant de détecter un pic potentiel (puisque celui-ci à déjà été identifié comme appartenant à la classe basse), ainsi on préfèrera avoir un modèle plus performant pour détecter les événements de la classe « haute » turbidité, quitte à accepter quelques faux positif supplémentaires. Ceci détermine le poids que l'on souhaite accorder à telle ou telle mauvaise classification. Bien évidemment il convient d'accorder plus d'importance aux évènements « haut » que « bas ».

Un format plus aisément interprétable veut que les résultats de classification qui sont sous forme matricielle, doivent être transformés en scalaires pour pouvoir mener des analyses de sensibilité en traçant la variation du critère de performance considéré en fonction des paramètres du réseau.

Afin de pouvoir comparer les modèles entre eux, le critère de performance doit pouvoir être indépendant des exemples qui le composent. Par exemple, si nous désirons comparer l'effet d'une répartition sur un modèle donné, il ne faut pas regarder les ensembles « Test » car ceux-ci ne sont pas formés des mêmes exemples donc ne sont pas comparables.

Format des résultats bruts

Matrice de classification

Chaque classificateur à un seuil donnera en résultat une matrice de classification (**MC**) 2x2 de la forme :

Tableau A-11 : Matrice de classification des résultats bruts

		Observé	
		Basse	Haute
Prédit	Basse	A	C
	Haute	D	B

(A, B, C, et D) étant 4 entiers : nombre d'exemples appartenant à chaque catégorie.

A et B sont des exemples bien classés, C est un faux négatif, D est un faux positif.

Trois ensembles

Chaque modèle de classification donnera une matrice de classification par ensemble d'échantillonnage (Train, Select, et Test). Train a servi pour l'apprentissage, il comprend de 60 à 80% des données, Select nous a servi pour optimiser un des paramètres de l'apprentissage (pour savoir quand arrêter afin d'éviter le sur-apprentissage) et Test représente des exemples non présentés au réseau lors de l'apprentissage. C'est la performance sur ce dernier qui est particulièrement intéressante pour évaluer la capacité de généralisation du modèle.

Un nombre d'exemples variables par saison et par seuil de turbidité

Selon la saison considérée et la valeur de turbidité seuil, nous n'avons pas le même nombre d'exemples disponibles dans les classes « basse » et « haute ».

Par conséquent, si les évènements « haut » ne représentent que quelques pourcents des exemples disponibles, leur mauvaise classification pourrait être masquée par de bons résultats dans la classe « basse ». Plus le seuil de turbidité augmente, moins le nombre d'exemples de « haute » turbidité disponibles est élevé.

Par exemple, nous avons pour l'automne :

Tableau A-12 : Nombre d'exemples observés par classe et par seuil - automne

Seuil	4 UTN	5,5 UTN	7,5 UTN	9,3 UTN
Turb obs basse	1231	1269	1294	1303
Turb obs haute	92	54	29	20

Il est intéressant de noter que le nombre d'exemples dans chaque classe va varier aussi selon la répartition d'échantillonnage considérée.

Définition des critères de performance

Afin de rencontrer tous les objectifs cités ci-dessus, et d'être plus sélectif sur le modèle optimal répondant tous nos besoins, nous avons opté pour une approche multicritères avec quatre indicateurs de performance :

1. Pourcentage de classification correcte TEST.
2. Matrice de perte TEST.
3. Matrice de perte SOMME.
4. Matrice de performance TEST.

À l'exception du troisième critère, tous ces indicateurs sont calculés à partir de la matrice de classification de l'ensemble de Test pour refléter la capacité de généralisation du modèle.

% classification correcte TEST

Il s'agit du premier critère directement donné par Statistica, mais aussi le plus parlant en termes de communication des résultats.

Il est calculé par la formule : $\frac{A+B}{(A+B+C+D)}$

Les résultats varient entre 0 et 1. Si notre classificateur était parfait, la valeur obtenue serait 1.

Cependant, il présente le gros désavantage, dans notre application, de considérer comme équivalent les événements « bas » et « haut ». Vu que nous souhaitons accorder plus d'importance aux pointes de haute turbidité, nous avons créé le deuxième critère.

Matrice de perte TEST

Souvent utilisé en classification, une matrice de perte (**MP**) attribuant un coût à chaque erreur dans les classes a été créée.

La matrice se présente sous la forme ci-après. Un événement bien classé ne « coûte » rien, alors qu'un événement mal classé se voit attribuer un coût variable.

Tableau A-13 : Matrice de perte

		Observé	
		Basse	Haute
Prédit	Basse	0	$2*Card(C_{basse\ ;5,5UTN}) / Card(C_{haute\ ;5,5UTN})$
	Haute	1	0

Où $(Card(C_{k\ ;5,5UTN}))_{k=\{basse;\ haute\}}$ représente le nombre total d'exemples (cardinal) de la classe C_k à la saison considérée et au seuil 5,5UTN. Le facteur 2 a été choisi arbitrairement, il a pour but d'accorder deux fois plus d'importance aux évènements « haut »,. Une étude économique sur les coûts d'opération de la station et sur le coût associé à une erreur pourrait définir précisément ce facteur; il faudrait cependant pouvoir donner un prix à la non prédiction d'un pic et à ces conséquences en termes de santé publique et d'opération (avis de bouillir, non respect des normes du RQEP), ceci est assez subjectif...

Nous visons le minimum de la fonction de coût associée qui s'écrit :

$$\sum_{j=1}^{2}\sum_{i=1}^{2}(ML)_{ij}.(MC)_{ij} \quad = \quad 1*D + 2*C*\frac{Card(C_{basse\ ;5,5UTN})}{Card(C_{haute\ ;5,5UTN})}$$

Le résultat est un réel positif ou nul, une classification parfaite donnant le résultat 0.

Ce critère nous donne une meilleure idée de la répartition des mauvaises classifications car il est plus sensible aux faux négatifs : ces derniers provoquant généralement une forte augmentation du score.

Matrice de perte SOMME

Il s'agit du même concept que précédemment mais appliqué à l'ensemble des exemples de la saison. Le nombre total d'exemples étant le même, quelque soit l'échantillonnage utilisé, sommer le score des fonctions de coût des matrices de perte

pour chaque ensemble (Train, Select, et Test), permet d'obtenir une idée de la performance globale du modèle et ceci autorise à comparer les répartitions entre elles.

Matrice de performance TEST

Ce critère vient combiner pourcentage de classification correcte et matrice de perte. Le principe est le même que pour cette dernière : nous créons une matrice attribuant des « coûts » à chaque situation et calculons le coût total engendré. Ici, la différence vient du fait que l'on donne un « crédit » aux bons classements, et un coût aux erreurs (valeur négative). La matrice de performance (**MP**) se présente sous la forme suivante :

Tableau A-14 : Matrice de performance, seuil x UTN

		Observé	
		Basse	Haute
Prédit	Basse	1	$-2*\text{Card}(C^{TEST}_{\text{basse, xUTN}}) / \text{Card}(C^{TEST}_{\text{haute , xUTN}})$
	Haute	-1	$2*\text{Card}(C^{TEST}_{\text{basse, xUTN}}) / \text{Card}(C^{TEST}_{\text{haute , xUTN}})$

Où $\text{Card}(C^{TEST}_{\text{basse, xUTN}})$ représente le nombre d'exemples « bas » de l'ensemble TEST au seuil x UTN. L'explication pour $\text{Card}(C^{TEST}_{\text{haute , xUTN}})$ est équivalente mais pour les exemples « haut ».

La fonction de coût associée s'écrit :

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(MP)_{ij}.(MC)_{ij}}{((A+C)*(MP)_{11}+(B+D).(MP)_{22})}$$

Le dénominateur étant un facteur de normalisation dans le cas où tous les exemples seraient bien classés. Le résultat obtenu est entre 0 et 1, une classification parfaite donnant la valeur 1.

Nous obtenons ici une version plus générale du pourcentage de classification correct car il prend en compte l'importance relative des évènements « haut » et attribue un malus aux mauvaises classifications, malus venant dégrader l'efficacité du modèle.

Toutefois, dans certains cas où ne disposant que de très peu d'exemples de haute turbidité, il se peut qu'il y ait plus de faux négatifs que d'événement bien classés (C > B), le score obtenu devenant ainsi négatif. Afin de limiter les scores négatifs (dans le cas où les malus négatifs soient plus forts que les crédits accordés à une classification correcte), nous retenons le rapport $MP_{22} = \min (\mathrm{Card}(C^{TEST}_{basse,\ xUTN}) / \mathrm{Card}(C^{TEST}_{haute\ ,\ xUTN}))$ minimal sur les répartitions considérées pour le seuil x UTN.

Tableau récapitulatif des matrices par seuil et par saison

Les chiffres nous ayant permis de calculer les scores de performance sont indiqués dans le tableau ci-après. Par souci de simplification, nous n'y avons écrit que le scalaire variant de chaque tableau, au lieu de la matrice complète.

Tableau A-15 : Tableau récapitulatif des matrices de perte et de performance par seuil et par saison – prédiction de TURB_DB

Saison	Matrice de perte, $(ML)_{12}$	Seuil de turbidité	Matrice de performance, $(MP)_{22}$
Automne	27	4 UTN	13
		5,5 UTN	27
		7,5 UTN	35
		9,3 UTN	84
Printemps	26	4 UTN	11,3
		5,5 UTN	26
		7,5 UTN	42
		9,3 UTN	62,8
Été	188	4 UTN	131